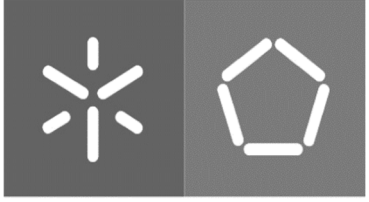
Universidade do Minho
Escola de Engenharia

Joana Catarina da Rocha Ferreira

Genomic and transcriptomic analyses in cancers related with viral infection

October 2016



Universidade do Minho
Escola de Engenharia

Joana Catarina da Rocha Ferreira

Genomic and transcriptomic analyses in cancers related with viral infection

Master's thesis

Master Degree in Bioinformatics

Work carried under the guidance of:

Luísa Pereira

(Supervisor)

Pedro Soares

(Co-supervisor)

October 2016

DECLARAÇÃO

Nome:

Joana Catarina da Rocha Ferreira

Endereço eletrónico: pg24094@alunos.uminho.pt Telefone: 918648001

Número do Bilhete de Identidade: 13925567 2 ZZ8

Título da dissertação:

Genomic and transcriptomic analyses in cancers related with viral infection

Orientador(es):

Luísa Pereira (Orientadora) e Pedro Soares (Co-orientador)

Ano de conclusão: 2016

Designação do Mestrado:

Bioinformática

Universidade do Minho, 19 / 10 / 2016

Assinatura:

AGRADECIMENTOS

A realização desta tese não teria sido possível sem a ajuda e constante incentivo dos meus orientadores, da minha família e amigos. Por todo esse apoio quero expressar minha mais profunda gratidão.

À Doutora Luísa Pereira por ter aceite ser minha orientadora, recebendo-me calorosamente no seu grupo e proporcionando-me uma experiência enriquecedora. Por toda a paciência e valiosas instruções e contribuições durante a execução deste estudo e na escrita desta dissertação. Quero também agradecer por todos os conselhos, apoio, disponibilidade e boa disposição durante este ano tão importante.

Ao professor Pedro Soares por ter aceite ser meu co-orientador desta tese e por ter sempre uma palavra simpática guardada para mim.

Às minhas colegas de grupo Andreia Brandão, Joana Pereira, Patrícia Marques, Sílvia Pereira, Susana Seixas e Verónica Fernandes pelo companheirismo, por serem sempre tão amigáveis e por se mostrarem tão prestáveis todas as vezes que pedia ajuda.

Ao Bruno Cavadas pela partilha de conhecimentos informáticos, pela incansável e constante ajuda, que se tornaram fundamentais para a execução desta tese. Assim como a orientação perante os mais variados obstáculos tendo sempre o cuidado de me dar o tempo e espaço necessários que me permitisse ultrapassá-los por mérito próprio.

À Marisa Oliveira, que embora seja o membro do grupo que à menos tempo conheço, se tornou numa das pessoas com a qual criei laços de amizade mais fortes. Agradeço-lhe todos os bons momentos, os constantes incentivos, ajudas e todo o resto (incluido as bolachas).

Aos meus colegas de mestrado Abel Sousa e João Silva por todos os momentos de descontração durante a hora de almoço.

À Diana Lemos pelos conselhos, trocas de ideias e opiniões, quer a nível profissional como pessoal. Mas, principalmente, por inconscientemente me transmitir força para nunca desistir.

Ao meu pai, Serafim Ferreira, que desde sempre foi o meu modelo de coragem. Sempre pronto para o que for preciso e para o que der e vier. O seu apoio incondicional, carinho e dedicação sempre foram os alicerces de todo o que até agora alcancei. O único defeito que tem (e que o persegue desde a juventude) é a falta de jeito para desenhar cadeiras.

Aos meus avós, Maria do Carmo Moreira e José Ferreira Júnior, pela constante preocupação, por todos os pequenos grandes mimos e por mostrarem todos os dias que as mais pequenas ajudas são das mais valiosas.

Por último, quero agradecer ao meu namorado e melhor amigo, Diogo Martins, por toda paciência, apoio, carinho e amor incondicional. Sendo ele a pessoa que me conseguiu mostrar em primeira mão que embora algo seja improvável, ou até mesmo quase impossível de acontecer, tal não quer dizer que nunca venha a aconteça e que nunca desistir de lutar por aquilo que acreditamos vale realmente a pena.

RESUMO

Nos últimos 30 anos foram-se acumulando evidências que têm vindo a apoiar a infecção viral como um factor responsável por 15-20% dos tumores malignos em humanos a nível mundial (W. S. Liang et al. 2014; McLaughlin-Drubin and Munger 2008). Estudos sobre os vírus oncogénicos demonstraram a sua importância no mau funcionamento celular ao longo do processo carcinogénico e demonstraram que a sua associação com o cancro varia entre 15% e 100% (McLaughlin-Drubin and Munger 2008), dependendo do tipo de tumor. Com a grande quantidade de informação genómica e metagenómica acessível nos consórcios internacionais públicos, tais como a base de dados TCGA, hoje em dia é possível inferir indiretamente infecções virais a partir de estudos genómicos centrados em humanos, uma vez que parte das *reads* irá alinhar com vírus e bactérias.

Tomando como ponto de partida a pesquisa feita por Tang et al. 2013, concentramo-nos nos cancros cervical (CESC), hepatocelular (LIHC) e da cabeça e pescoço (HNSC), que são conhecidos por apresentar uma alta proporção de casos virais-positivos (Tang et al. 2013). Fizemos *download* de dados RNA-Seq de 309, 424 e 566 amostras, respectivamente, e comparamos *unmapped reads* contra uma base de dados viral de referência (retirada da base de dados do NCBI) usando as ferramentas Batch, SAMTOOLS, bowtie e PRINTSEQ. A quantificação de cada vírus foi feita usando partes por milhão (ppm) e apenas vírus com ppm acima de 10 foram considerados como estando a infectar positivamente uma amostra. Confirmamos que cerca de 94% das amostras de CESC foram infectadas, principalmente por HPV (papilomavírus humano) e, especificamente, pela estirpe HPV16. Quase 32% das amostras LIHC foram infectadas por HBV (vírus da hepatite B). E por volta de 17% de amostras HNSC foram infectadas e o HPV16 foi o vírus mais comum.

A avaliação de enriquecimento diferencial de vias metabólicas entre grupos infectados e não infectados, para cada tipo de cancro, foi realizada por GSEA. Os sinais de enriquecimento para infecção e vias relacionadas com sistema imune eram evidentes no grupo infectado CESC, enquanto nos grupos infectados de LIHC e HNSC o enriquecimento era principalmente relacionado com replicação e reparação de DNA. Este facto parece indicar que a infecção é especialmente ativa no CESC, contradizendo alegações anteriores de que a tumorigenese no colo do útero não estava diretamente ligada à infecção. Nos três tipos de cancro, os vírus integraram os seus genomas no genoma do hospedeiro, afetando a replicação, manutenção e reparação do DNA. No nosso estudo sobre a integração do genoma de HPV16 numa amostra

de tumor HNSC, foi confirmada a integração viral no gene humano *RAD51B* que codifica uma proteína implicada na reparação de DNA por recombinação homóloga. Desta forma, conseguimos confirmar que HPV16 pode atuar tanto como agente cancerígeno directo e indirecto.

Provavelmente através da integração do genoma viral no genoma do hospedeiro, a infecção aumentou a quantidade de mutações somáticas no grupo de amostras infectadas em LIHC, mas não em HNSC onde o consumo de tabaco é também um importante agente cancerígeno. O reduzido número de amostras não-infectadas em CESC não permitiu uma comparação fiável da quantidade de mutações somáticas entre grupos de infectados e não-infectados. Ainda assim, nos grupos infectados de LIHC e HNSC, algumas mutações somáticas ocorreram no contexto de vias relacionadas com o sistema imunológico, mostrando que podem contribuir para tornar estes indivíduos susceptíveis à infecção.

Além disso, ao verificar a expressão dos genes de HPV16 em cinco amostras de CESC e de HNSC, confirmou-se que os genes E6 e E7 estão entre os mais expressos em muitas das amostras, enquanto que o E2 não é expresso. Os genes E6 e E7 são conhecidos por serem preferencialmente integrados no genoma do hospedeiro, ao contrário do gene E2, o qual controla a expressão daqueles, que não é integrado ou é fragmentado. Acredita-se que é a sobre-expressão de E6 e E7 que inicia a carcinogénese.

As taxas de infecção viral inferidas neste trabalho por *mining* de bases de dados omicos são muito semelhantes aos obtidos pelos métodos tradicionais (Tang et al. 2013), mostrando que a informação disponível nos consórcios internacionais públicos pode elucidar, indiretamente, sobre o envolvimento da infecção viral na tumorigénese. O elevado número de amostras por tumor, a grande variedade de origem geográfica das amostras e a caracterização de alto rendimento para diferentes plataformas omicas permitem comparações e avaliações múltiplas, numa escala não acessível anteriormente.

PALAVRAS-CHAVE

Cancro, infecção viral, RNAseq, Carcinoma do colo do útero, Carcinoma hepatocelular, Carcinoma da cabeça e pescoço.

ABSTRACT

In the past 30 years, accumulated evidence has been supporting viral infection as one factor responsible for 15-20% of human malignancies worldwide (W. S. Liang et al. 2014; McLaughlin-Drubin and Munger 2008). Studies on oncogenic viruses have proved their importance on cellular malfunction along the carcinogenic process, and showed that their association with cancer can amount from 15% to 100% (McLaughlin-Drubin and Munger 2008), depending on the type of tumour. With the large amount of genomic and metagenomic information available on public international consortia, such as TCGA database, it is nowadays possible to indirectly infer viral infections from the human centred omics studies, as a portion of the reads will align in viruses and bacteria.

Taking as starting point the research made by Tang et al. 2013, we focused on cervical (CESC), hepatocellular (LIHC) and head and neck squamous cell (HNSC) carcinomas, which are known to show a high proportion of viral-positive cases (Tang et al. 2013). We downloaded RNAseq data from 309, 424 and 566 samples, respectively, and run the unmapped reads against a reference database of viruses (downloaded from NCBI) by using the tools Batch, SAMTOOLS, Bowtie and PRINTSEQ. Quantification of each virus was performed using parts per million reads (ppm) and only viruses with ppm above 10 were considered as positively infecting the sample. We confirmed that around 94% of CESC samples were infected, mostly by HPV (Human papillomavirus) and specifically by the HPV16 strain. Nearly 32% of LIHC were infected by HBV (hepatitis B virus). Almost 17% of HNSC samples were infected, and the HPV16 was the most common present virus.

The evaluation of differential enrichment of metabolic pathways between infected and non-infected groups, for each cancer type, was performed in GSEA. Signs of enrichment for infection and immune related pathways were evident in CESC infected group, while in LIHC and HNSC infected groups the enrichment was mostly related with DNA replication and repair. This seems to indicate that infection is especially active in CESC, contradicting previous claims that tumorigenesis in cervix was not directly linked with infection. For the three cancer types, the viruses integrate their genome in the host genome, affecting DNA replication, maintenance and repair. In our investigation of integration of HPV16 genome in one HNSC tumor sample, we confirmed integration in the human *RAD51B* gene that codes a protein involved in DNA repair by homologous recombination. We thus confirmed that HPV16 can act both as indirect and direct carcinogen.

The infection, most probably through the integration of the viral genome in the host genome, increased the amount of somatic mutations in the infected group in LIHC, but not in HNSC where tobacco consumption is also an important carcinogen. The low number of non-infected samples in CESC did not allow a reliable evaluation of changes in the amount of somatic mutations. Even so, in both LIHC and HNSC infected groups, some somatic mutations occurred in the context of immune-related pathways, showing that they can contribute to render these individuals susceptible to infection.

Also, when checking expression of HPV16 genes in five samples each from CESC and HNSC, we confirmed that E6 and E7 genes are amongst the ones more expressed in many samples, while E2 is not expressed. E6 and E7 have been said to be preferentially integrated in the host genome, while E2, which controls their expression, is not integrated or it is disrupted. It is believed that the overexpression of E6 and E7 initiates carcinogenesis.

The viral infection rates inferred here from mining the omics databases are very similar to the ones evaluated by standard methods (Tang et al. 2013), showing that public international consortia can indirectly provide interesting insights into the involvement of viral infection in tumorigenesis. The high number of samples per tumor, the wide geographic origin of the samples, and the high-throughput characterisation for different omics platforms allows multilayer comparisons and evaluations, in a scale not affordable before.

KEYWORDS

Cancer, Viral Infection, RNAseq, Cervical carcinoma, Hepatocellular carcinoma, Head and neck squamous cell carcinoma.

INDEX

FIGURES INDEX

TABLES INDEX

ANNEXES TABLES INDEX

LIST OF ACRONYMS

BAM – Binary Alignment/Map format

BC – Bonferroni correction

BCR – Biospecimen Core Resource

CDS – coding sequences

CESC – Cervical carcinoma

CGHub – NCI's Cancer Genomics Hub secure repository

DCC – Data Coordinated Centers

DNAseq – DNA sequencing

EBV – Epstein Barr virus

ES – Enrichment Score

FDR – False Discovery Rate

GCC – Genome Characterization Centers

GDAC – Genome Data Analysis Centers

GO – Gene Ontology

GO_BP – Gene Ontology Biological Process

GO_CC – Gene Ontology Cellular Component

GO_MF – Gene Ontology Molecular Function

GSC – Genome Sequencing Centers

GSEA – Gene Set Enrichment Analysis

HAdV – Adeno-associated virus

HBV – hepatitis B virus

HCV – hepatitis C virus

HHV – herpesvirus

HIV – The human immunodeficiency virus

HNSC – Head and neck squamous cell carcinoma

HPV – Human papillomavirus

HTS – high-throughput sequencing

ICGC – The International Cancer Genome Consortium

LDL-R – low-density lipoprotein receptor

LD – lipids droplets

LIHC – Hepatocellular carcinoma

MCV – molluscum contagiosum virus

miRNAseq – MicroRNA sequencing

NAFLD – non-alcoholic fatty liver disease

NCI – the National Cancer Institute

NES – Normalized Enrichment Score

NGS – next generation sequencing

NHGRI – National Human Genome Research Institute

NIH – US National Institutes of Health

NT – normal tissue samples

PC – principal components

PCA – principal component analysis

PLAT – Percutaneous local ablative therapy

ppm – parts per million

RFA – radiofrequency ablation

RNAseq – RNA sequencing

RPPA – Reverse-phase protein array

SAM – Sequence Alignment/Map format

TCGA – The Cancer Genome Atlas

TP – tumour tissue samples

TSS – Tissue Source Sites

1. INTRODUTION

In the past, the occurrence of "house cancers" started the idea that cancer could be caused via some infectious agents. These "house cancers" were known to occur in people living in the same place, mostly between married couples and transmitted from mother to child. In the 19th century, these observations encouraged studies that could test the idea of development of cancer malignancies through bacteria, fungi or parasites. Initially, these studies were unsuccessful in finding such agents, leading to a step-back in the hypothesis for many years (McLaughlin-Drubin and Munger 2008).

Some significant researches revealed cancer formation with cell-free transmission in non-malignant tumours in animals, but not in human models (McLaughlin-Drubin and Munger 2008). This was the case of Francis Peyton Rous experiments in 1911 that showed that the sarcomatous chest tumour on chicken could be transmitted over cell-free tumour excerpts, confirming once more the hypothesis of the involvement of small infectious agents. Like many before him, these studies were treated like curiosities for many years. Only in the 1950s, after Ludwik Gross proved that murine retrovirus and polyomavirus induced murine cancers, Rous research was fully appreciated and awarded a Nobel Prize in 1966 (McLaughlin-Drubin and Munger 2008; Moore and Chang 2010).

Following the success in proving viral infection in animal cancers, a crescent number of scientists started a demand to discover oncogenic viruses in humans. Yet, it would take more 53 years since Rous's famous research for the first human oncogenic virus to be observed. In 1964, during a study in cell lines of Burkitt's lymphoma from African pediatric patients, Anthony Epstein, Bert Achong and Yvonne Barr identified small particles on the electron microscope. These particles were then identified as a virus by virological studies and named Epstein Barr virus (EBV) in their honour (Morales-Sánchez and Fuentes-Pananá 2014; Moore and Chang 2010). Later on, in 1970, hepatitis B virus (HBV) was discovered by D. S. Dane in human cells by using hepatitis B surface antigen (McLaughlin-Drubin and Munger 2008). Thanks to these two studies and to the development of model systems, additional studies on viral infection in cancer led to the identification of seven more oncogenic viruses, suggesting that more viruses important in cancer development can still be discovered (McLaughlin-Drubin and Munger 2008; Morales-Sánchez and Fuentes-Pananá 2014).

1.1. Viral Infection in cancer

Surprisingly, only a small portion of people infected with oncogenic viruses actually develop cancer, and an even smaller part of that group of people transmits the infection. Cancers seem to be a final event of the viral infection, and eventually lead to host death and viruses destruction (Moore and Chang 2010; Morales-Sánchez and Fuentes-Pananá 2014). Tumour viruses usually maintain their persistence in the host by creating a chronic infection during a period of many years, with a very small production, or almost none, of viral particles (Morales-Sánchez and Fuentes-Pananá 2014).

Studies through the past century led to the conclusion that infectious agents (viruses, bacteria and parasites) should be divided in two groups in relation to their involvement in cancer: direct and indirect carcinogens.

Direct carcinogen agents can be found in the cancer cells in a monoclonal form, which indicates that these agents are responsible for tumour cell transformations by keeping the carcinogen phenotype through expressing at least one transcript. This oncogenic persistence is maintained by viral formation of episomes or incorporation into the host genome (Morales-Sánchez and Fuentes-Pananá 2014; Moore and Chang 2010). In summary, direct carcinogen viruses have three essential characteristics: the viral genome can be found in every tumour cell; once host cells have grown, the virus can immortalize; and the virus causes cell transformation, immortalization and migration by disturbing the cell normal functioning (Chen et al. 2014). Human papillomavirus (HPV), EBV and molluscum contagiosum (MCV) are examples where this activity is easily seen (Figure 1A) (Morales-Sánchez and Fuentes-Pananá 2014; Moore and Chang 2010; Chen et al. 2014).

On the other hand, indirect carcinogen agents lead to cancer through infection and inflammation that eventually causes cell mutation, showing that these agents are not restricted to remain inside the host cells (Morales-Sánchez and Fuentes-Pananá 2014; Moore and Chang 2010). Thus, these viruses work over two principal ways: starting chronic inflammation and immunosuppression. Chronic inflammation is initiated by chemokines produced by mutated cells that attack immune cells and injure local tissue. In this way, this kind of tumour activity resumes to a cycle between infections, inflammations and local tissue destruction (Figure 1B). In a similar manner, the immunosuppression process stars with an immune response failure.

The most common example of this process is displayed by HIV virus, which infection is responsible for lowering the host defences and consequently increasing the risk of new infections to occur in the host (Figure 1C) (Morales-Sánchez and Fuentes-Pananá 2014).

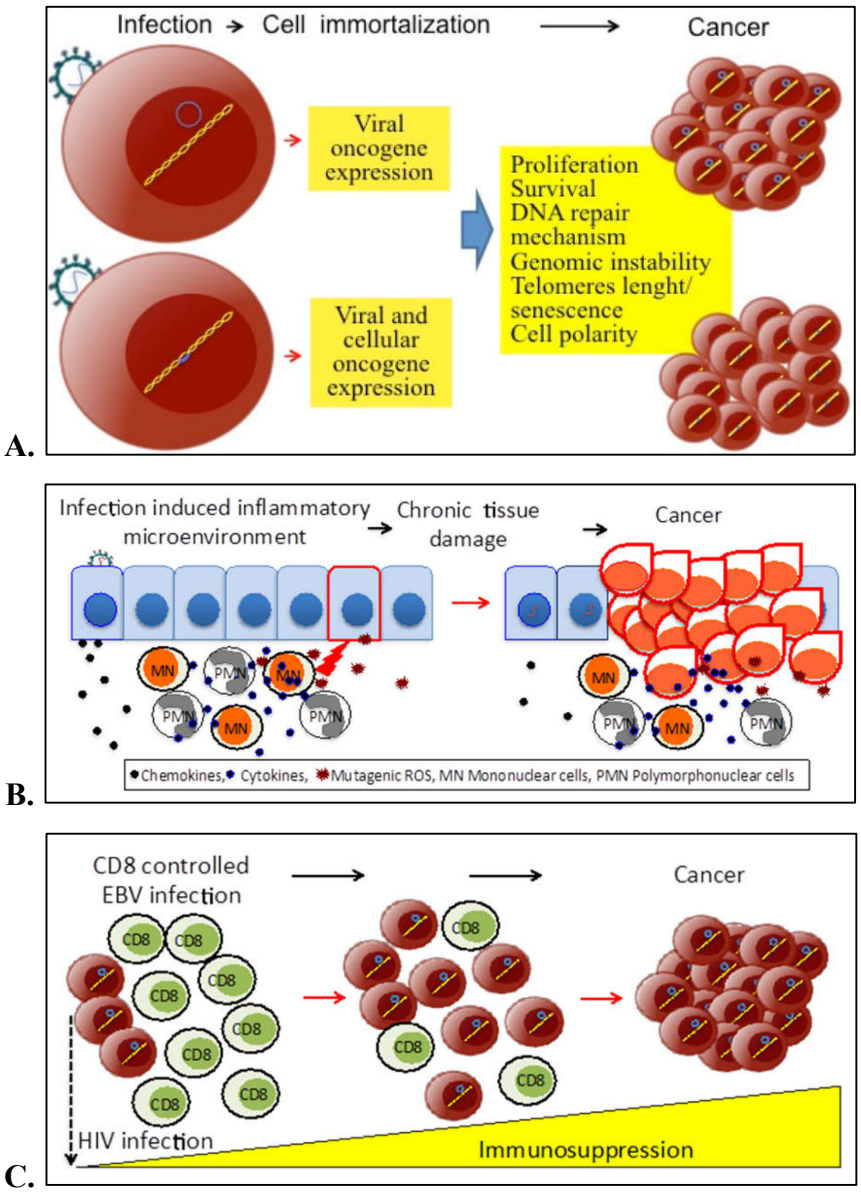


Figure 1. Direct and indirect viral carcinogenesis processes (images taken from Morales-Sánchez et al. 2014). **(A)** Representation of direct viral carcinogenesis. Superior Section: Formation of episomes by viral genomes (e.g. herpesvirus). Lower Section: Viral integration into de host DNA (e.g. retroviruses). **(B)** Representation of chronic inflammation of indirect viral carcinogenesis. Production of chemokines from infected cells which attack immune cells and damage the local tissue. **(C)** Representation of immunosuppression of indirect viral carcinogenesis. Immunosuppression is caused by HIV and EBV infection, and is controlled by cytotoxic CD8 T cells. While HIV infection develops, immune system starts failing and the host becomes more venerable to EBV infection (Morales-Sánchez and Fuentes-Pananá 2014).

Still, there are many viruses that cannot be classified only into direct and indirect mechanisms, such as HBV and HCV (hepatitis C virus). HBV genome is integrated into the host cells in almost all HBV-related tumours, however it is still not clear if the virus requires cell proliferation to maintain the genome transcription or not. Nevertheless, this type of classification is

still important, as it is a practical way to classify cancers that have a bigger chance of occurring via the action of oncogene viruses (Morales-Sánchez and Fuentes-Pananá 2014; Moore and Chang 2010).

Some studies have recently showed new techniques for viral detection on high-throughput DNA and RNA sequencing. These new studies allow an efficient unbiased detection of viruses in extended collections of data from cancer samples, overcoming limitations of studies based on low-throughput methodologies. One such study was performed by Tang et al. (2013), whom mapped virus infection in 4,433 samples from 19 human cancers, by using human centred RNASeq data. The inference based on RNASeq guarantees that the viral genome is being transcribed and that the infection is active. From the results obtained by Tang study it is possible to acknowledge that cervical (CESC), hepatocellular (LIHC) and head and neck squamous cell (HNSC) carcinomas are the top three cancers with higher viral infection rates (96.6%, 32.4% and 14.8%, respectively) (Tang et al. 2013). However, this work was based in a limited number of the samples now available in TCGA: 87 of 309 for CESC (28%); 34 of 424 for LIHC (8%); and 304 of 566 for HNSC (54%).

1.1.1. Cervical carcinoma (CESC)

Nowadays, the second principal cause of female death is cervical cancer. Thanks to large endorsements for vaccination and constant control through Papanicolaou test (Pap test), the frequency of this carcinoma dropped drastically in developed countries, but not yet in developing countries, where 80% of cases occur (Adams et al. 2014; Muñoz et al. 2003). HPV has been reported in closely 100% of cervical cancers, mainly the high risk HPV16 and HPV18 strains. This infection can occur from sexual contact or via vertical transmission from mother to child during pregnancy or during birth (Adams et al. 2014).

Even if many women are infected with HPV, in most cases the immune cells eliminate the infection agents, and only a small portion of infected woman will progress from infection to cervical cancer. These women normally manifest immunity deficiencies caused by inherent genomic instability or because of bad life style habits, like smoking (Adams et al. 2014; Canavan and Doshi 2000).

1.1.2. Hepatocellular carcinoma (LIHC)

Hepatocellular carcinoma is one of the most common cancers worldwide and one of the most rapid cause of death. This cancer is common in Asia and Africa, but recently it is also

rising in America and Europe (Siegel and Zhu 2009; Chen et al. 2006). The principal risk factors for LIHC are known to be abusive alcohol consumption and infection through HBV and HCV; however, it has been reported that 5-30% of the patients are cryptogenic (unknown disease cause). The majority of LIHC cases are associated to a metabolic syndrome manifested in the liver named non-alcoholic fatty liver disease (NAFLD), consisting in fat deposits in liver cells in absence of any history of excess alcohol consumption. NAFLD is also related with insulin resistance, obesity, diabetes, dyslipidemia, and other metabolic conditions (Siegel and Zhu 2009).

The best treatment to this cancer is still the partial hepatectomy, though it is only applicable to about 9-27% patients. Percutaneous local ablative therapy (PLAT) and radiofrequency ablation (RFA) are being applied and allowing a better cancer control (Chen et al. 2006).

1.1.3. Head and neck squamous cell carcinoma (HNSC)

This cancer occurs 90% of the times from squamous cell carcinomas in mucosal surfaces in the neck and head. The principal risk factors have been reported to be dangerous life style habits, like ultraviolet light exposure and tobacco and alcohol consumption. Recently it was proved that HPV infection is a major cause, mostly in oropharyngeal cancer (tonsillar and base of tongue cancer). Till recently, 66% to 95% of the cases occurred in men after 50 years of age, but with the increasing number of female smokers this number tends to vary (Adams et al. 2014; Egloff et al. 2014).

Patients diagnosed with this cancer and HPV positive showed improvement and higher survival rates when treated with chemotherapy, radiation and surgery, than the HPV negative counterparts. This led to the idea that HPV condition is a good biomarker of HNSC. As no diagnostic tools are available, HNSC patients must check for persistent symptoms such as aching throats, inflamed glands and oral wounds (Adams et al. 2014).

1.2. Lipids influence in viral carcinogenesis infection

Lipids are the most important components in cell membranes, playing important roles in many biological functions like cell growth and division, energy production, motioning, maintenance of cell membrane integrity, activity of membrane enzymes, and DNA helix stabilization (Raju et al. 2014). Cholesterol is a lipid that executes some of those functions, and the lipoprotein receptors placed in the cell surface are responsible for its absorption and control within cells. This mechanism is very important during membrane fusion for entry of virions in the

cells and during particle maturation. Taking the example of non-enveloped viruses, its capsid interacts with the host membrane and usually controls some of the membrane constituents, like glycosphingolipids, to start the viral infection. Some viruses, like HCV, can even induce viral entrance with functional low-density lipoprotein receptor (LDL-R) (Heaton and Randall 2011).

Likewise, lipids play an important role during the viral life cycle since this occurs in the membranous cell organelles (as endoplasmic reticulum), assisting the membrane curvature and drafting core proteins to enable the construction of virus particles (Heaton and Randall 2011; Konan and Sanchez-Felipe 2014). Other example of its importance is the recognition of lipids droplets (LDs) as sites for assembly of some viruses (e.g. HCV). LDs are organelles derived from endoplasmic reticulum which are abundant in cholesterol esters and triglycerides (Konan and Sanchez-Felipe 2014). Furthermore, there are some changes in lipid metabolism and in specific lipid species, which are expected to be important in viral replication, assembly and secretion events. Some of those changes are the increase of flux metabolic directed into fatty acid biosynthesis pathway (visible in metabolic analysis of cytomegalovirus infection), and the inhibition of cholesterol and sphingolipids biosynthesis, or knockdown of proteins involved in their transport (visible during HCV production) (Heaton and Randall 2011; Konan and Sanchez-Felipe 2014).

In the other hand, oncogenic studies have showed that cholesterol is found in low concentration and can indicate a continuous oncogenic process, in some growing tissues and blood partitions (Singh et al. 2013). So far, most claims of variations in lipid profiles among normal and tumour cells (like in CESC and HNSC) (Raju et al. 2014; Singh et al. 2013) have been associated with bad life habits like the use of tobacco, and not with infection. It has been reported that tobacco carcinogens are responsible for greater amount of peroxidation of polysaturaded fatty acids through induction of free radicals and production of reactive oxygen species. This lipid peroxidation can heavily disturb fundamental membrane cell components and may lead to carcinogenesis since, because of it, great utilization of lipids like cholesterol, lipoproteins and triglycerides is needed for membrane syntheses of the new cells. In order to provide for these needs, lipids are taken from blood circulation, degraded from lipoproteins or metabolised.

The elucidation of the involvement of lipids in cancer is very important, namely in terms of treatment. Some studies have shown that antioxidant vitamins have a great defensive influence against lipid peroxidation (Patel et al. 2004). It was also suggested that cancer cells or several minor malfunctions during the lipid or antioxidant vitamin metabolism can be responsible for

the drop of lipid concentration that occurs during hypolipidemia (Raju et al. 2014; Singh et al. 2013; Patel et al. 2004). Yet, these associations are controversial since many studies have reported a different relation between cancer and lipid profile parameters (Raju et al. 2014; Patel et al. 2004). It is also not yet guaranteed that cancer diagnosis through lipid parameters is viable (Raju et al. 2014).

1.3. Database and bioinformatic tools

Following the publication of the human genome sequence, an increase in large-scale genome studies has taken place, namely in the cancer research field. Soon after, the idea of a unified record of cancer genomes emerged in order to deal with the following issues: independent tumour genome studies could lead to investment of resources in the same type of cancer instead of broadening the spectre of types of cancer studied; the absence of normalisation between the different studies could produce noise and biases in the analyses and comparisons; the existence of many cancers that differ through the world should lead to a coordinated strategy between local consortia; and the fact that only one organisation could coordinate and, consequently, accelerate the distribution of data and a more efficient analysis (Hudson et al. 2010).

Researchers and funding agency councils from 22 countries, encouraged from that initial idea, gathered in late 2007 in Toronto, Canada, to decide the best way to proceed. The International Cancer Genome Consortium (ICGC) was launched, aiming to perform a comprehensive and high-throughput analysis of tumours in an organized way (Zhang et al. 2011). Given the dimension of this challenge, operational groups were formed in order to create policies and plans that would form the base requirements to participate in the ICGC (Hudson et al. 2010).

In order to hasten cancer research, the consortium has as main goals: (1) the creation of complete sets of genomic anomalies in tumours of 50 distinct cancer types and subtypes, in high resolution, fullness, high quality and controlled data; (2) form complementary lists of transcriptomic and epigenomic datasets from the correspondent tumours; and (3) allow the whole investigation community to have access to the data obtained promptly with the least limitations. Thus, ICGC also organized the research community so that no repeated researches occur, and promoted the proliferation of new technologies, software and techniques (Hudson et al. 2010; Zhang et al. 2011).

The large amount of genetic information available in these databases raises bioethical issues, particularly concerning the patient anonymity. Since the genomic data allow individual identification, data access policies were developed in order to protect that information. Those access policies consist in dividing the dataset into two different sets (Table 1). The first set will contain accessible information and will not have data capable of identifying the individual identity. The second set will be the most restricted group, and although it will not contain direct identification of the individual, it will have complex genomic and clinical data that belong to a unique human being. The permission of access to the second set can only be given by the Data Access Compliance Office to researchers that submit their project (Hudson et al. 2010; Zhang et al. 2011).

Table 1. *List of sets of information from each group of access (Zhang et al. 2011).*

ICGC Open Access Datasets	ICGC Controlled Access Datasets
→ Cancer Pathology – Histologic type or subtype – Histologic nuclear grade → Patient/Person – Gender – Age range → Gene Expression (normalized) → DNA methylation → Genotype frequencies → Computed Copy Number and Loss of Heterozygosity → Newly discovered somatic variants	→ Detailed Phenotype and Outcome Data – Patient demography – Risk factors – Examination – Surgery – Drugs – Radiation – Sample – Slide – Specific histological features – Protocol – Analyte – Aliquot → Gene Expression (probe-level data) → Raw genotype calls → Gene-sample identifier links → Genome sequence files

Initially, the consortium started by including two European consortia and 10 participating countries, but currently the number increased to 15 (Hudson et al. 2010). Another important project that contributed to ICGC development was The Cancer Genome Atlas (TCGA). This project is older than ICGC, and it was initiated by the US National Institutes of Health (NIH). These two related projects were the main factors of the great growth of complete comprehension of cancer genomes (Hudson et al. 2010).

1.3.1. TCGA

The Cancer Genome Atlas is a public funded project by the US National Institutes of Health and was created as a Pilot Project from the Nacional Institute of Health (NIH) in 2006, with similar aims to ICGC but at a local scale in the US. It is currently using large-scale genome sequencing and multi-dimensional analyses in order to characterise above 30 human cancers (Hudson et al. 2010; Tomczak et al. 2015).

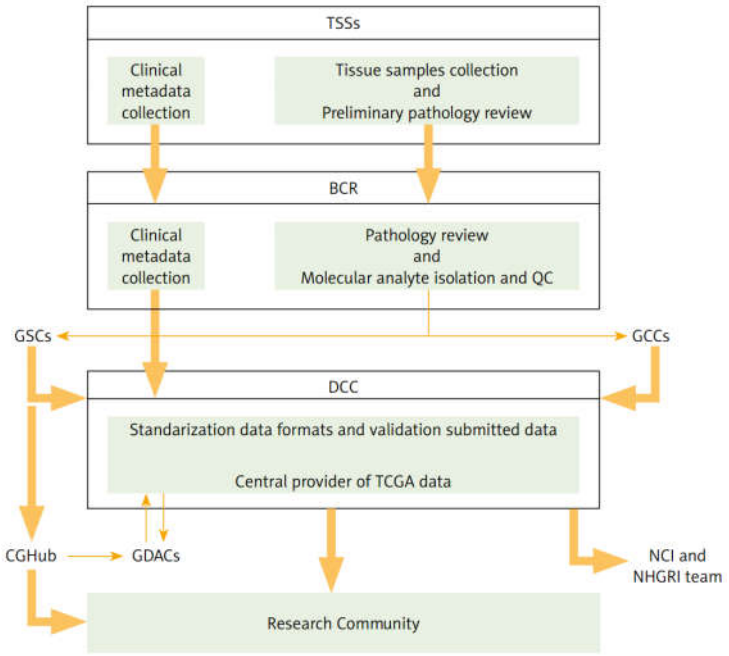


Figure 2. TCGA structure and relation between partners. TSSs (Tissue Source Sites) is responsible for clinical metadata and biospecimen assemble from authorised cancer patients. BCR (Biospecimen Core Resource) then approves these data collected from TSSs. After approbal, processing and validation of the quality and quantity of sample, BCR registers and submits metadata to DCCs (Data Coordinated Centers). At the same time, molecular analysis are provided to GCCs (Genome Characterization Centers) and GSCs (Genome Sequencing Centers) for additional genomic characterization and high-throughput sequencing. Sequence-related data is stored in DCC. GSCs submits trace files, sequences, and alignment mappings to CGHub (NCI's Cancer Genomics Hub secure repository). Then genomic data submitted to DCC and CGHub are made accessible to the research community and to GDACs (Genome Data Analysis Centers). GDACs supply new information-processing analysis and visualisation tools to the research community (Tomczak et al. 2015).

TCGA is well structured and divided in many partners who are responsible for different tasks, from the sample gathering and manipulation, to high performance methods based on microarrays and next-generation sequencing, and respective bioinformatic data analysis (Tissue Source Sites, Genome Characterization Centers, Genome Data Analysis Centers, etc.). For a better understanding of the cancer omics, different approaches have been applied: RNA sequencing (RNAseq), MicroRNA sequencing (miRNAseq), DNA sequencing (DNAseq), SNP-based platforms, Array-based DNA methylation sequencing, and Reverse-phase protein array (RPPA) (Tomczak et al. 2015).

NIH, the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) established partnerships with other North American and European institutes, and in 2010 TCGA data was largely incorporated the ICGC Data Portal (Tomczak et al. 2015). Yet, TCGA is not allowed to officially become ICGC's member for legal and technical manners, once it belongs to the US's NIH (Figure 2) (Hudson et al. 2010).

1.3.2. Bioinformatic tools

With the growing number of available genomic data, the demand on fast and efficient analyses led to the need of developing new bioinformatics tools. Many of the recently implemented tools were used by Tang et al in 2013 that, in fact, proved their applicability. Some of the studies performed in this thesis are based on the pipeline implemented by Tang and his colleagues.

1.3.2.1. SAMtools

The first step on the processing of data is the alignment for a competent read mapping against a reference sequence. Diverse alignment tools were available but generated different formats that made complex the downstream processing, and rendering it urgent to develop an universal format. This led to the genesis of the Sequence Alignment/Map (SAM) format (Figure 3) (Li et al. 2009).

The SAM format is simple and flexible, once it can retain information from numerous sequencing platforms and read aligners, and supports single and paired-ended reads and blending reads of various types. This format is prepared to support alignment sets over 1011 base pairs which is approximately the size needed for a human genome. A similar format is the Binary Alignment/Map (BAM) format, which consists in a binary representation of the SAM format. Basically it has the same information of a SAM file but is represented in a more compacted way that can process information from a particular location in the genome without loading the entire file data, in order to improve performance (Li et al. 2009). Nowadays, almost all genomic studies use SAM/BAM files as the base files for diverse types of analyses. But many of the current informatic methods take too much RAM memory and require several CPUs, so it is very important to use the correct and more appropriate tools for each research (Langmead et al. 2009).

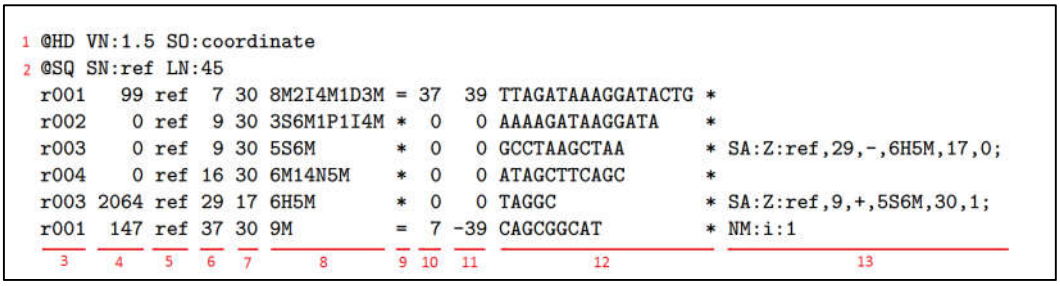


Figure 3. Representation of a SAM file. **1-** Header Line. VN indicates the format version and SO the sorting order alignment. **2-** Sequence dictionary. SN represents de sequence name and LN the sequence length. **3-** QNAME, Query template NAME. **4-** FLAG, bitwise FLAG. **5-** RNAME, Reference sequence NAME. **6-** POS, 1-based left end mapping POSition. **7-** MAPQ, MAPping Quality. **8-** CIGAR, CIGAR string. **9-** RNEXT, Reference name of the mate/NEXT read. **10-** PNEXT, Position of the mate/NEXT read. **11-** TLEN, observed Template LENgth. **12-** SEQ, segment SEQuence. **13-** QUAL, ASCII of phred-scaled base QUALity+33 (Li et al. 2009).

SAMtools consists in a collection of software packages which par and manipulate the existent alignments in the SAM/BAM files and have the capability to: (1) sort and combine alignments; (2) exchange other alignment formats to SAM or BAM format; (3) eliminate PCR replicas; (4) call SNPs and short indel variants; (5) create an output with information listed per-position; and (6) show this output in a text-based format (FASTA) (Li et al. 2009).

The only disadvantage of this software is the fact that it takes a little too long for alignment and indexing of data. For example, to use a MAQ file with 112GB it takes about 10h to convert it and 40 min to index it with less than 30 MB of memory. This tool has implementations using two different languages, Java and C, which grant different functionalities. Since SAMtools is such an efficient tool, for the past six years it has become a crucial instrument for next generation sequencing (NGS) alignment studies (Li et al. 2009).

1.3.2.2. PRINSEQ

PRINSEQ is a bioinformatic tool that is used to filter, reformat and trim NGS data, and that is used to create a pre-processing statistic synopsis of data and of its quality, permitting an efficient control before any downstream study being made (Schmieder and Edwards 2011). This tool supports FASTA, FASTQ and QUAL files of genomic and metagenomic datasets, and it was implemented in two different ways: (1) in a web interface, which is very easy to use; and (2) a standalone lite version (Schmieder and Edwards 2011). The principal difference between these two versions is that the standalone version does not create any statistic synopsis in graphical form.

1.3.2.3. Bowtie

Bowtie was developed to create an ultrafast, memory-efficient short read aligner optimized for mammalian re-sequencing. Bowtie can align short DNA sequence reads to long sequences in a fast way and using an acceptable amount of RAM memory. By being based in full-text minute-space index, it applies Burrows-Wheeler indexing which results in a memory occupation of 1.3GB for a human genome. With this low RAM memory occupation, it allows an ordinary computer, with 2GB RAM, to use this software.

Bowtie was especially optimized to achieve the highest performance. Because of that if more than a match exists for the same read, it is assured that one will be reported, but not necessarily the best alignment. In other words, Bowtie may fail valid alignments if there are many mismatched reads. To avoid this, Bowtie has an option that increases the precision but it will reduce its performance (Langmead et al. 2009). This tool is very practical, and supports FASTA and FASTQ files, and Bowtie2 supports also insertions, deletions or paired-end alignments (Langmead et al. 2009; Langmead and Salzberg 2012).

1.3.2.4. Picard Tools

Set of bioinformatics tools for high-throughput sequencing data manipulation compatible with SAM, BAM, CRAM and VCF files. All Picard Tools only work through command line and are provided as a single executable jar file (Java file) (Broad Institute 2016).

Even though there are around 80 tools available for a high variety of tasks, in this thesis we only used two of those tools – SamToFastq and MarkDuplicates. SamToFastq has as primary objective to convert SAM or BAM files into FASTQ. Furthermore, it also has options which allow to extract read sequences and base quality scores from the input files. In this thesis this tool will be mostly used to split whole genome files into forward and reverse sequences that will serve as input files in the bioinformatic tool VirusSeq (which will be explained below) in order to detect viral integration in human genomes (Broad Institute 2016).

MarkDuplicates locates and tags duplicated reads in BAM and SAM files. Then, these duplicate reads are defined as originated from the same DNA fragment and acknowledged as read pairs which have identical 5' positions for both reads in mate pair. As an output, a new SAM or BAM file is created with all duplicated sequences removed, identified in the SAM flag field, or even marked with a duplicate type in the "DT" optional attribute. Additionally, it also outputs a second file with the number of read pairs examined, unmapped reads, unpaired reads, duplicated unpaired reads, duplicated read pairs and optical duplicated read pairs (duplicates

that appear clustered together during sequencing and can emerge from optical/image-processing artifacts or from bio-chemical processes during clonal amplification and sequencing) (Broad Institute 2016).

1.3.2.5. VirusSeq (with Mosaik)

As it was said before in this report, many viruses are related to human tumours and with the improvement of bioinformatic technologies it became possible to detect viral presence/infection in human cells through the analyses of paired-end reads by VirusSeq. This algorithm uses NGS data and detects known viruses and its integration sites in the human genome.

This tool uses FASTQ format files with paired-end reads as input, and those reads can be whole-transcriptomic or whole-genomic data. The first step is the alignment to the reference genome, by using the program Mosaik (provided when VirusSeq is downloaded). This program is an aligner tool that, contrary to what happens with Bowtie, can compare at the same time the viral genome against the tumour genome in study so that the viral integration can be determined. VirusSeq starts the identification of human and non-human sequences by aligning the input file with the human genome reference. Then, the non-human sequences are compared against the viral database in order to find any virus present in those sequences. The genome sequences of well-known viruses are all put together into one single chromosome named chrVirus, which will serve as a viral database. This viral database was combined with the human reference genome hg19 making this reference genome named hg19Vrus where the chrVirus was designed as chr25 (Chen et al. 2013). With this strategy ViruSeq can, then, identify if any of the viruses is infecting the inputted sample and describe its precise location when inserted in the human genome.

VirusSeq is a very effective tool, yet, it also has some crucial limitations: (1) the virus database must be updated frequently by the user (Chen et al. 2013); (2) it takes too much memory RAM, and an ordinary computer cannot run it; (3) and it takes a long time to run.

1.3.2.6. PCA

The big datasets being produced and analysed currently have a high amount of variables and variance contributing to it. This is the case of the expression profiles for the around 20,000 human genes that is inferred from RNAseq data. Several parameters can contribute with variability to the human gene profile, besides the cancer and infection status, such as the date when the lab analysis was perform, the lab that performed it, the geographical origin of the sample,

etc. Some theoretical techniques allow to investigate the partitioned effect of the multifactorial variables, as the older and commonly used principal component analysis (PCA) (Jolliffe and Cadima 2016). PCA is a classical method which reduces de data dimension by transforming a new set of variables called principal components (PC) which are unrelated and orthogonal. PCs are ordered so that the kth PC have the kth higher variances of all PCs. As PCA can detect artefacts and biases, it is usually used as pre-processing phase (Y. Liang et al. 2005; Yao et al. 2012).

However, this technique is not perfect. It may fail when it comes to reproduce biology relations. Recent studies have showed that microarray gene expression may have a super-Gaussian distribution, and the PCA method considers that gene expression follow a multivariate distribution. Also, PCA reduces the data based on the maximization of its variance, and sometimes biological problems do not relay on the data highest variance (Yao et al. 2012).

1.3.2.7. Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a computational process that evaluates differences among two biological states of a set of genes and determines if they are statistically significant (Subramanian et al. 2007). Its principal goal is to determine if a defined set of genes S are randomly distributed through a list L of mRNA expression profiles of genes ordered according to their differences. The GSEA method works on three principal steps: (1) calculation of an Enrichment Score (ES) that, shows the degree of representation of set S in the extremes of the entire ranked list L (top or bottom); (2) estimation of Significance Level of ES by using an empirical phenotype-based permutation test that conserves the correlation arrangement of the gene expression data; and (3) adjustment of estimated significance level to consider for multiple hypothesis testing (Subramanian et al. 2005). GSEA also offers the possibility to perform preranked analysis, where the gene set enrichment analysis is run against the ranked list of genes created during the default GSEA run. This allows to perform a correction of results from the default GSEA run, and to obtain more accurate values.

Two indicators are important in GSEA results, the False Discovery Rate (FDR) and the Normalized Enrichment Score (NES). FDR is the probability of a gene set represents a false positive; it is a ratio of the actual ES vs. the ES of all gene sets against all permutations of the dataset and the actual ES vs. ES of all gene sets against the actual dataset. On the other hand, the NES value represents, as its name indicates, the normalization of the ES value and can be used to compare analysis results all over the gene set.

There are many tools that execute similar functions as GSEA, but it is beyond doubt that GSEA continues to be the most used. Many have reported lack sensitivity of this method, however many improvements have been made do correct that limitation. The fact that GSEA was conceived to find general dissimilarities in a cumulative distribution probably gave advantage of GSEA over other methods like z-test, which only could identify sets of genes with mean shifts (Irizarry et al. 2009).

1.3.2.8. G:Cocoa (G:Profiler)

It is known that most of the somatic mutations in cancer are usually unique events occurring independently in different patients, with the exception of the ones affecting oncogenes and proto-oncogenes (Lawrence et al. 2015). But these independent somatic mutations can affect the same metabolic pathways, and in our specific case, make infected individuals more susceptible to the viral infection. To test this we used G:Cocoa from G:Profiler webtool (Reimand et al. 2016).

G:Profiler is a public web server developed and maintained by the Institute of Computer Science of University of Tartu with the objective of characterising and manipulating gene lists of high-throughput genomics. This web server is currently available for over 80 species and has six available tools, as G:Cocoa tool used in this work (Reimand et al. 2016).

G:Cocoa is a enrichment tool which performs functional analysis of dozens of gene lists allowing a simple and minimal view of functional enrichments of those lists and providing a means to rank and compare gene lists through their functional annotations. Because of the multiple functional enrichment analysis and its comparison against many reference lists, this tool performs a multiple testing correction which methodically reduces the significance of each p-value, discarding the number of false positive values. By default it is used the tailor-made algorithm g:SCS, otherwise the user may choose to use Bonferroni correction (BC) or Benajamini-Hochberg False Discovery Rate (Reimand et al. 2016).

1.3.2.9. HTSeq

HTSeq is a Python package which allows the fast processing and analysis of high-throughput sequencing (HTS) data. This tool receives as input the most common file formats for reference sequences, short reads, short read alignments, genomic feature, annotation and score data (FASTA, FASTQ, SAM/BAM, GFF, VCF, etc.). The principal component of this tool is

a container class which allows working with genomic coordinates (genomic positions or genomic intervals) more simply. This tool contains two different applications: HTSeq-qa, for quality assessment; and HTSeq-count, for processing RNA-seq alignments for expression analysis (Anders et al. 2015).

HTSeq-count receives as input SAM/BAM files from RNAseq data and GTF/GFF files with gene models. The number of times each gene had its exons overlapped in the aligned reads is counted. It is important to notice that ambiguous reads (reads which overlap with more than one gene or that align to various positions) are not considered. Even if this tool is written in Python, it can be run with simple shell commands without any Python knowledge. Also, this tool requires very little RAM memory, being able to process over 1.2 million reads per minute using around 250 to 450MB RAM, approximately (Anders et al. 2015).

2. AIMS

In order to better understand the impact and effect of viral infection in cancer, this study has six principal aims:

1) to estimate infection rates in CESC, LIHC and HNSC from inference on the TCGA available RNASeq data. The work of Tang et al. 2013 quantified in 425 samples belonging to these cancer types, but we enlarged to the full amount of 1299 samples currently available at TCGA. The lower quantity of cases checked in Tang et al. work will be used as quality control in our analysis;

2) to fully characterize the virus strains that are present in each infected individual, by using an updated human viral reference database, according to information deposited in NCBI;

3) to check if specific human pathways are associated with the infection, by performing gene enrichment analysis (using GSEA tool) between infected and non-infected groups in the three types of cancer;

4) to determine if infection is responsible for an increasing of somatic mutations, or if these are related with immune-related pathways;

5) to identify which viral genes are expressed during infection of the most prevalent viruses;

6) to detect viral insertion within a human genome.

3. METHODS

3.1. Viral database

For the construction of the reference viral database, a list of accession numbers of all known human viruses was downloaded through the NCBI complete viral genome website. Then, all bacteriophages and endogenous viruses were removed from the list, making a final total of 5003 viruses. A Python script was created, to receive that list as input. Then, by using the Bio package provided freely by NCBI that contained the module Entrez and their efetch function, the genomes of all viruses in that list were downloaded into a single fastq file. This final fastq file was then used as reference database to investigate viral presence in the cancer samples.

3.2. Human raw RNAseq database

Since our intention was to investigate active viral infection, implying transcription of viral genes, we downloaded the dataset of raw RNAseq samples for each cancer. This dataset aimed to address the human transcriptome, but other non-human transcripts are also indirectly screened, namely viral transcripts.

The RNAseq data was downloaded from the CGHUB database, provided by TCGA, filtering the data only for RNAseq with HG19 assembly, accounting to 309 samples from CESC cancer, 424 from LIHC and 566 samples from HNSC. All RNAseq samples were downloaded as BAM files and each sample had an average size of 7GB which corresponded to a total size of 8.62TB of downloaded information (2.14TB for CESC, 2.49TB for LIHC and 3.99TB for HNSC). In mean, around one week was needed for the download for each cancer type. Most of these RNAseq samples were from the solid tumor, but a few were from the solid normal tissue. We extracted both in order to do some comparisons.

3.3. Viral presence detection

Once all samples had been downloaded, the viral presence procedure could be implemented. Initially all BAM files were indexed through SAMTOOLS which allows to analyse directly specific parts of the BAM file without reading through all the sequence. In this case, SAMTOOLS allowed us to focus on the unmapped regions against the known human transcriptome, information previously processed by TCGA team. These unmapped regions were extracted from the BAM files and converted to fastq files. A quality control was then applied

to the fastq files, by using PRINTSEQ program, which removes reads with less than 45 nucleotides.

Thereafter, Bowtie tool was used to align each sample (the cleaned fastq files of unmapped reads) against the viral database previously created, and limiting the data to 2 mismatches and a maximum of 25 valid alignments for each read, outputting the alignment in a SAM file format.

Finally, we used SAMTOOLS and Picard Tools to remove duplicated reads from the SAM file and count how many times each virus was aligned within each sample. The final output is a text file with a list of viruses found in each sample and the number of times each virus aligned with that sample. This whole process took around 4 to 5 days to be completed for each cancer type analysed here in a dual-core computer with 4GB RAM memory.

3.4. Infection state determination in a cancer sample

Following others (Tang et al. 2013), we calculated the parts per million (ppm) value for each virus in each sample. In order to do that the number of each virus reads on that sample was divided by the total number of reads in the same sample and multiplied by one million, according to the following formula:

$$ppm = \frac{Virus\ reads}{Sample's\ total\ reads} \times 10^6$$

The total number of reads in the sample is the sum of all human and non-human reads which were previously determined using SAMTOOLS. We report all viruses present in a sample, even with low ppm, but we established a threshold for a sample to be considered as infected: when the most prevalent virus is >10 ppm.

3.5. Human transcriptomic profile and matrix construction

In order to investigate human gene expression differences between infected and non-infected groups, we used the gene expression values inferred by TCGA team and deposited in their database. Single files for each sample, with the complete list of normalized expression counts for all known human genes (20,531 genes), were downloaded (average of 400MB each) and merged into one matrix per cancer type. That matrix had in the first column the list of all genes and in the successive columns the normalized counts for each sample. A matrix with this

distribution was used in the GSEA software that we will discuss bellow, while a reverse one was used in the PCA analyses.

3.6. PCA

Gene expression is influenced by several factors that can lead to spurious or false results in statistical analyses of differential expression. In order to have an idea on the variance among the samples we performed principal component analyses (PCA) (Mardia et al. 1979) by using an R script. We controlled for variance between tumour and normal tissue samples, and between infected and non-infected samples.

3.7. Gene Expression Analysis

For a gene expression enrichment evaluation between infected and non-infected groups, we performed a GSEA analysis (Subramanian et al. 2007). Both groups were constructed in order to have an equivalent number of individuals, controlled by the group with the lower number of samples and the other made of the same size by picking up samples at random controlling for the population group and geographical origin to which the individuals belonged. In the end, size of groups amounted to 18 in CESC, 118 in LIHC and 86 in HNSC.

This tool analyses if significantly differentially expressed genes are organised within known metabolic pathways, so it incorporates reference lists of metabolic pathways from diverse datasets as KEGG (Ogata et al. 1999) and Gene Ontology (Consortium 2001). We performed the analyses against KEGG and the three main Gene Ontology lists (Biological Process, Molecular Function and Cellular Component).

When running GSEA we used most of the default parameters, except for a small number of exceptions. First we started by changing the option “Collapse dataset to gene symbols” from True to False, since our matrix already had gene names there is no need to associate to any chip annotation file, and by setting the gene “Max size” and “Min size” parameters to 1000 and 5, respectively. This option allowed us to exclude from the analyses metabolic pathways having a too limited number of genes.

Once each GSEA run had finished we used the same data to do a Preranked run. We left almost all parameters as default only changing the “Collapse dataset to gene symbols”, the gene “Max size” and “Min size” parameters like we did previously on the normal GSEA run.

In the results, pathways with false discovery rate (FDR) lower than 0.25 are considered as significantly differentially expressed between the two groups, and they are ranked based on the Normalized Enrichment Score (NES). We present the top of the metabolic pathways in each cancer type in a graphical form by using R.

3.8. Somatic mutations in infected and non-infected groups

In order to test the hypothesis if viral infection leads to increased somatic mutation, we downloaded the curated somatic mutation profiles reported by TCGA team in their website. Each file presented all mutations, their functional implications and respective genes and samples. We counted number of mutations per class of type of mutation in infected and non-infected groups, and calculated the ratios of these mutations (by dividing number of mutations per number of samples). The results are shown through bar plots performed through R.

To test if independent somatic mutations affect the same metabolic pathways we used G:Cocoa from G:Profiler webtool (Reimand et al. 2016) The curated files of somatic mutations were used to count the number of times each gene was hit in infected and non-infected groups from each cancer. Then, those genes wore sorted in descending order through those counts, divided by infected or non-infected groups and used as input in G:Cocoa tool.

In this tool the inputted information is then compared with information available in the same four references list of metabolic pathways used in GSEA (KEGG, the Gene Ontology Biological Process, Molecular Function and Cellular Component reference lists). Then, the results are presented in a list of metabolic pathways significantly mutated in each or in both groups. Bar plots (representing –log(p-value)) for each infected and non-infected groups were built in R for the most significant pathways.

3.9. Viral integration on the host genome

With viral presence confirmed in the cancer samples we could now investigate its integration into the host genome, the most direct proof of an active infection and potentially providing additional information of contribution to cancer development (depending on the integration location in the human genome). We began by using a HNSC sample known to have viral integration (Tang et al. 2013), in order to check if we were able to correctly detect it. For this we

used VirusSeq tool, which uses Hg19 as reference to human genome and has an extra chromosome which corresponds to the reference viral database from which each tested sample would have to be compared.

We started by downloading the whole genome bam file from the TCGA database (total size of 200GB). Then, by using the Picard Tool SamToFasq, we divided the sample into two, corresponding to the forward and reverse sequences, and converted it into fastq files. These two fastq files will serve, then, as input files for VirusSeq, which was run with all default options.

Unfortunately, because this tool uses a large amount of RAM memory and takes a very long time to finish the whole process (approximately two weeks), it must be run in a powerful server. Since we had limited access to a server we could only perform the investigation of viral integration in one sample.

3.10. Expressed Viral genes

We also identified the most expressed viral genes for the most frequent viruses in each cancer type, by using HTSeq toll (ref). We downloaded a GFF file with information of all coding sequences (CDS) for HPV16 and HBV from NCBI database and then five samples per cancer where those viruses hade a higher infection rate. GFF and SAM files from each sample that were a result output after aligning each sample with the viral database were used as input in HTSeq. Each run took around 2 hours in the same computer used for viral presence detection, since the most time consuming part was already performed during the previous viral infection detection.

As output it is then presented the number of times each gene was expressed in every sample as well as information about ambiguous reads, not aligned reads or even low quality reads. Samples from different cancers that were infected by the same virus, CESC and HNSC, were compared.

4. RESULTS AND DISCUSSION

4.1. Cervical carcinoma (CESC)

In the total 309 samples from the CESC cancer downloaded from TCGA, 304 samples were from tumour tissue (TP) with 286 infected and 18 not infected (94.1% infection), 3 samples from normal tissue (NT) with 1 infected and 2 not infected, and 2 metastic samples, one infected and another not infected (Figure 4; Table 1 in annexes).

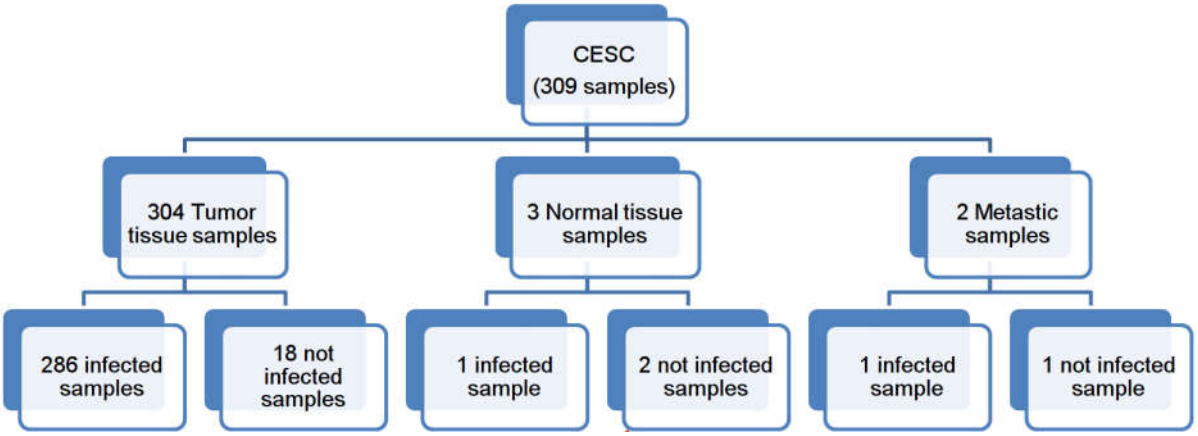


Figure 4. *RNAseq samples distribution through tissue type and infection results in samples from CESC cancer.*

We were limited to three patients for which we had both normal and cancer tissue samples, and only one had infection in both tissues, while two had only infection in the tumor (at least according to our classification as infected, with ppm>10).

The results of the rate of infection in CESC (Figure 5) meet the expected outcome of nearly 100% infected cases for this type of tumour (Tang et al. 2013). The predominant virus in the 94.1% infected CESC samples is HPV (Figure 6), and diverse strains are observed. However, the more frequent one across samples is HPV16 and this is also the one that reaches extremely high ppm values (above 300ppm). High values of ppm are also relatively common in HPV18 and HPV45 that infect a similar number of patients (approximately 22% each) (Figure 6). It is also important to mention that HPV31 is a very common virus among the infected samples too. This virus infects only a few samples (6%) with ppm above 300, however, it infects a very large number of samples (51%) with ppm between 10 and 300 ppm. Likewise, even though not as common as the previous viruses, HPV26 is present in a considerable number of samples. Some of those samples have a considerable high value of ppm (between 50 and 300; 13%) yet, the big majority have ppm from 10 to 50 (16%). In the same line of thought, HPV33, HPV58

and HPV68b are present in a very small number of samples where these viruses have a very high ppm, like 100 to 300 or above 300, there are a considerable number of samples where they are present with ppm between 10 and 50 (Figure 6). HHV are very rare (mostly present in ppm between 2 and 10) and present in a small number of samples (in total, 41 samples).

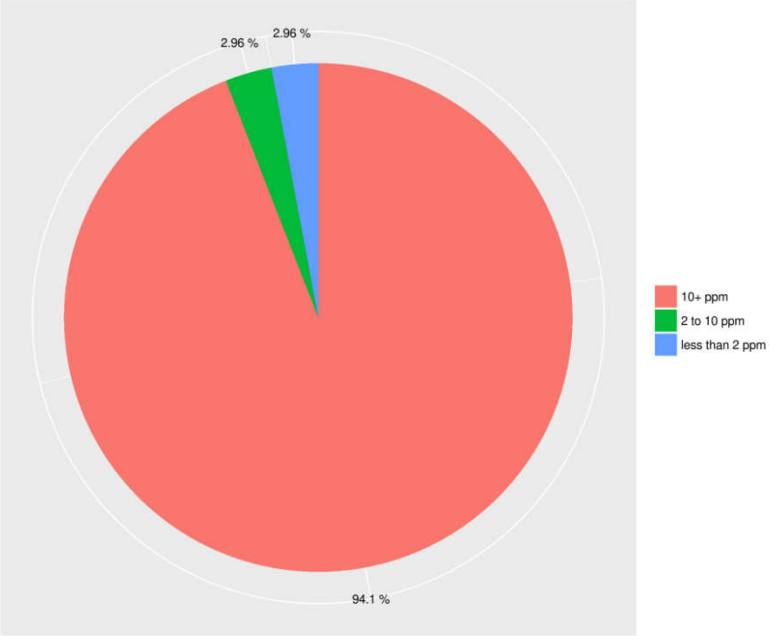


Figure 5. *Percentage distribution of viral infection in CESC samples. Each sample was represented by the virus with the higher ppm and only tumour tissue samples were considered in this graphic. Virus with ppm lower than 10 were not considered as infecting a sample.*

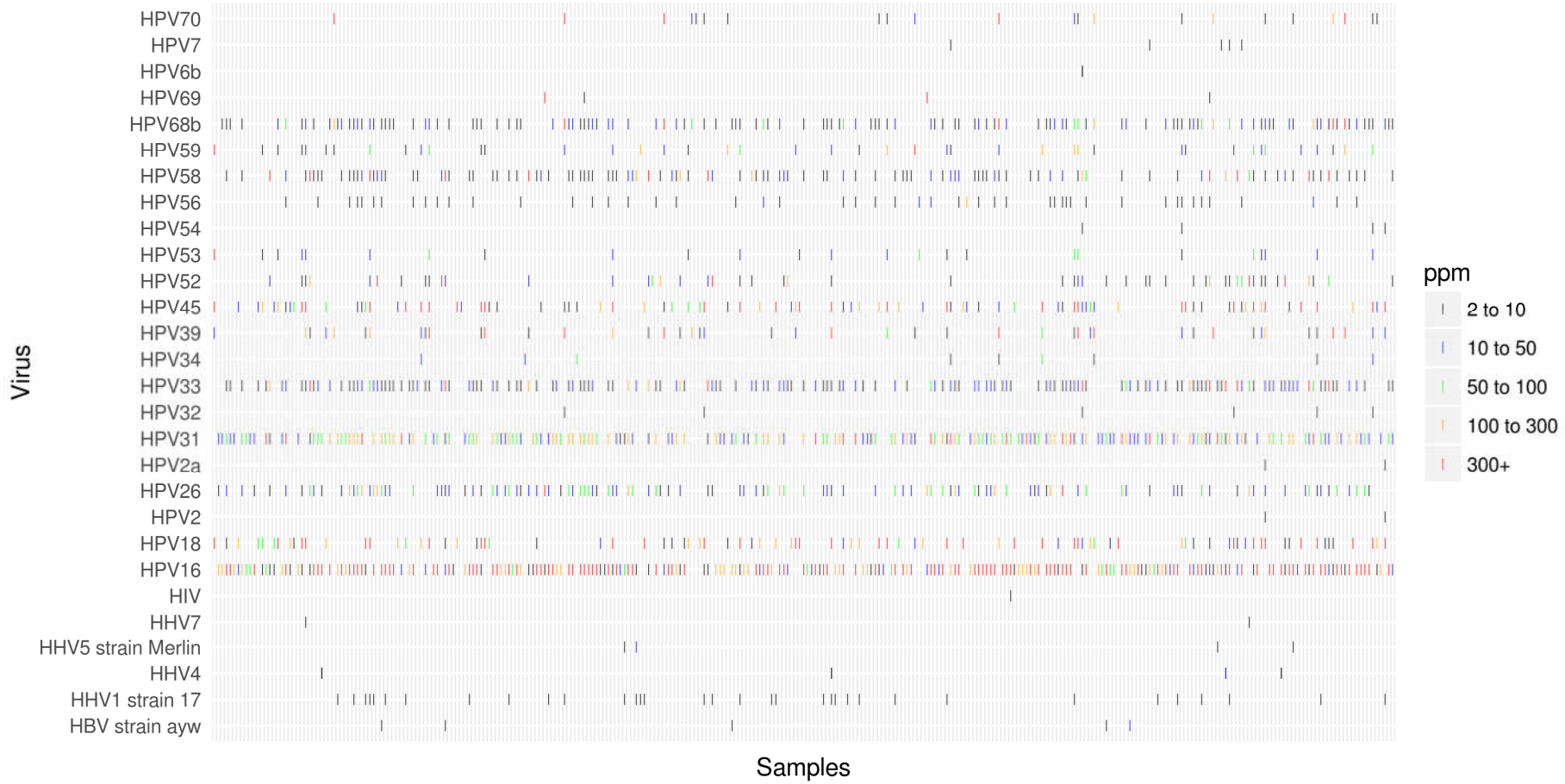
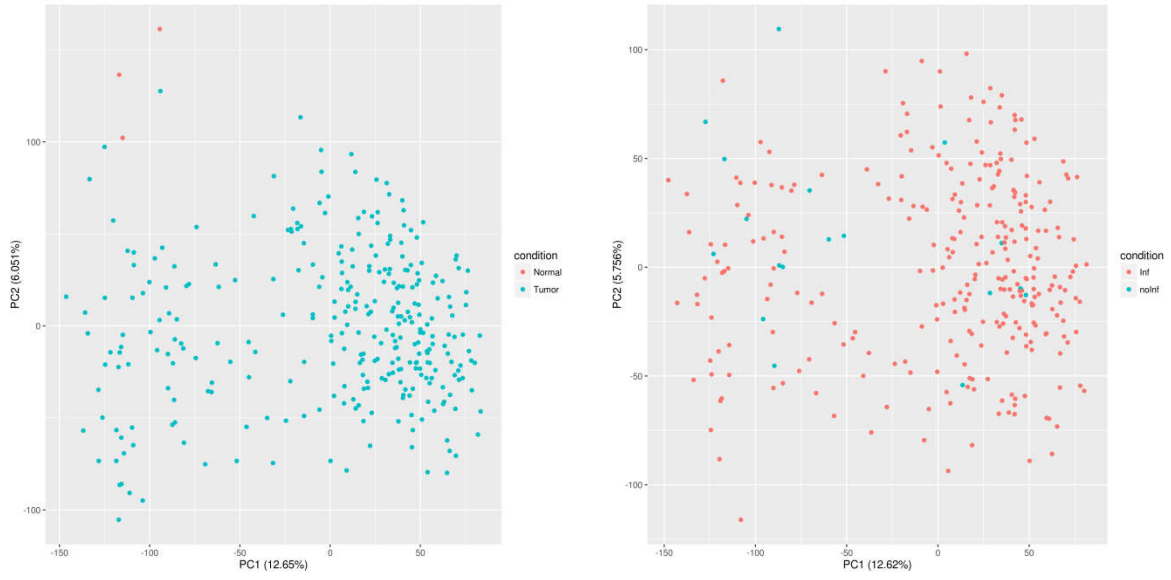


Figure 6. *Viral presence distribution in CESC cancer samples by ppm. Each column represents one sample. Viruses with ppm smaller than 2 were not considered.*

We checked the effect of variance in the gene expression profiles through two PCAs. In the first one (Figure 7A), we checked the effect of TP versus NT samples. The NT samples, even though in a very small number, are distanced from most TP samples, showing that there are significant differences in gene expression between the normal and tumor tissues. But even more interesting, the TP samples are organised in two clusters, which deserves further investigation. In the second PCA (Figure 7B), only the TP samples were considered, to check the variability between infected and non-infected groups. As we can see, the split is not perfect, being 12 non-infected samples dispersed in the cluster of the left, and the remaining six in the other cluster. The low number of non-infected samples precludes the application of a statistical test.



A B

Figure 7. *PCA of CESC samples. (A) Comparing samples between tumour and normal tissues. (B) Comparing infected and not infected TP samples.*

We then identified all tumor infected samples present in the two clusters, concluding that there were 61 and 225 samples on the left and right cluster, respectively. After, we checked their associated clinical information in order to detect any relevant factor, and could not see any significant difference which could explain the two clusters. We then evaluated if the clustering could be due to a different profile of virus infecting the samples (Table 2). When considering only strains with ppm>10, there is a bigger percentage of HPV18 and HPV45, a considerable higher percentage of HPV33, HPV39 and HPV53 and a smaller percentage of HPV31 infection on the left cluster in comparison with the right cluster. While HPV16 and HPV52 are more frequent in the right than the left cluster. So, it seems that differences in proportions of

the HPV strains that infect the patients can contribute to differential expression profiles of the human genome.

Table 2. *Count of samples bearing infection in each cluster of the PCA of Figure 7B. The percentage was calculated using the number of samples infected by each virus with ppm>10 and the total of infected samples in each cluster (61 samples in the left cluster and 225 in the right cluster).*

Virus	Left Cluster		Right Cluster	
	Nº of samples	%	Nº of samples	%
HPV16	36	59.02	154	69.68
HPV18	31	50.82	36	16.29
HPV26	20	32.79	69	31.22
HPV31	29	47.54	147	66.52
HPV33	22	36.07	46	20.81
HPV39	13	21.31	16	7.24
HPV45	32	52.46	36	16.29
HPV52	4	6.56	21	9.50
HPV53	10	16.39	6	2.71
HPV56	1	1.64	4	1.81
HPV58	8	13.11	32	14.48
HPV59	12	19.67	17	7.69
HPV68b	17	27.87	34	15.38
HPV34	0	0	5	2.26
HPV70	0	0	11	4.98
HPV4	0	0	1	0.45
HPV5	0	0	1	0.45
HPV69	0	0	2	0.90
HHV4	0	0	1	0.45
HBV	0	0	1	0.45

The main genes responsible for the variability in PC1, in decreasing order, were: *KRT14*, *DSG3*, *CLCA2*, *DSC3*, *SERPINB13*, *CALML3*, *SPRR1B*, *TMPRSS11D*, *KRT6C*, *BNC1*, *KRT6A*, *RHCG*, *IVL*, *MUC5B*, *PKP1*, *SPRR1A*, *KRT5*, *PROM1*, *SPRR2A*, *LASS3*, *TP63*, *SBSN*, *GBP6*, *SPRR2D* and *KRT6B*.

GSEA was used to evaluate differences in expression profiles between the infected and non-infected groups (Figure 8; genes involved in some pathways are displayed in Table 2 in annexes). In the bar plot for GO_BP, it is immediately visible a few obvious pathways related to viral infection in top of the graphic (e.g.: cellular defence response, defence response, immune response, response to virus, inflammatory response, immune system process, endosome transport, JAK-STAT cascade, response to other organism, T-cell activation), and even the pathways related to cell death are also important in the defence to viral infection.

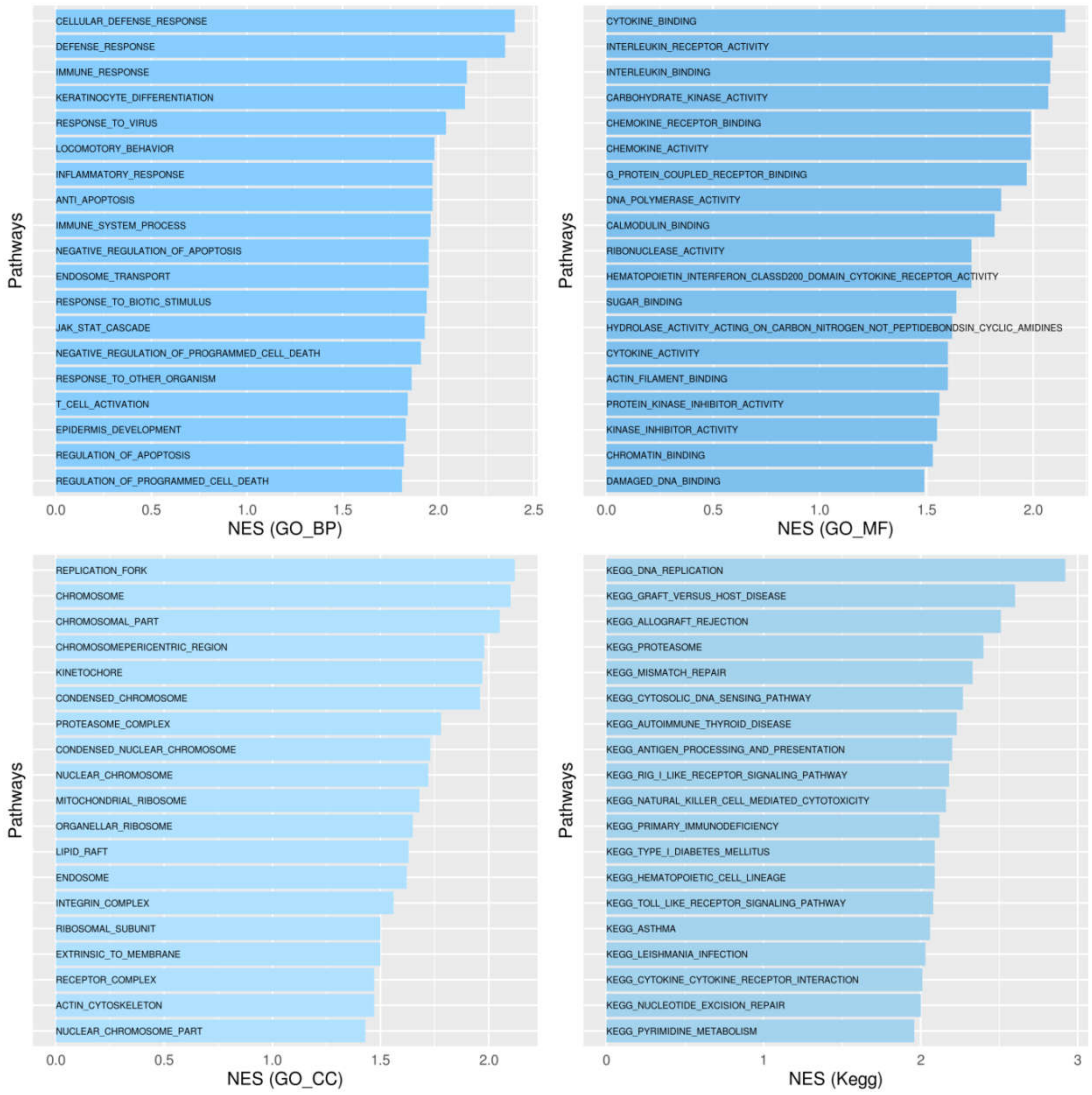


Figure 8. *GSEA results for CESC when comparing infected and non-infected samples and using the Gene Ontology Biological Process (GO_BP), Molecular Function (GO_MF) and Cellular Component (GO_CC), and the KEGG references lists. For each graphic 19 metabolic pathways with FDR smaller than 0.25 were picked and then sorted by NES value. The positive NES values mean that these are more expressed in the infected than in the non-infected.*

Considering the GO_MF bar plot, all pathways involving interleukins, cytokines and chemokines are related with the immune system. In fact, interleukins and chemokines are types of cytokines, a group of small proteins that are important in cell signalling and being produced by diverse types of cells, including immune cells like macrophages, B lymphocytes, T lymphocytes and mast cells. Besides these, pathways related to DNA repair were also overexpressed in the infected group.

When analysing the GSEA GO_CC, pathways related with chromosomes are at the top of differentiation, concordant with an active viral infection and integration of viral in the host genome. But also pathways related to membrane movement and lipid rafts are highlighted, being more related with entrance and replication of the virus in the cell.

Finally, in the results using the KEGG reference list, a mix of the metabolic events seen in the previous GO lists is observed: some relevant pathways directly related to infection like proteosome regulation, natural killer cell mediated cytotoxicity, primary immunodeficiency, antigen processing and preservation; other pathways related to DNA replication, mismatch repair, cytosolic DNA sensing pathway; and diseases related with infection or inflammation, as graph versus host disease, type I diabetes, asthma and leishmanial infection.

We also downloaded the curated file of somatic mutations in CESC, but given that this only contained 194 samples of which only 10 were from non-infected patients, we decided to not follow with analyses on the possible influence of infection in the total number of somatic mutations.

4.2. Hepatocellular carcinoma (LIHC)

From the 424 RNAseq samples available in TCGA for LIHC (Figure 9; Table 3 in annexes), 371 were from TP with 253 non-infected and 118 infected (infection rate of 31.8%), 50 from NT with 40 non-infected and 10 infected, and 3 samples from recurrent tumour tissue all infected.

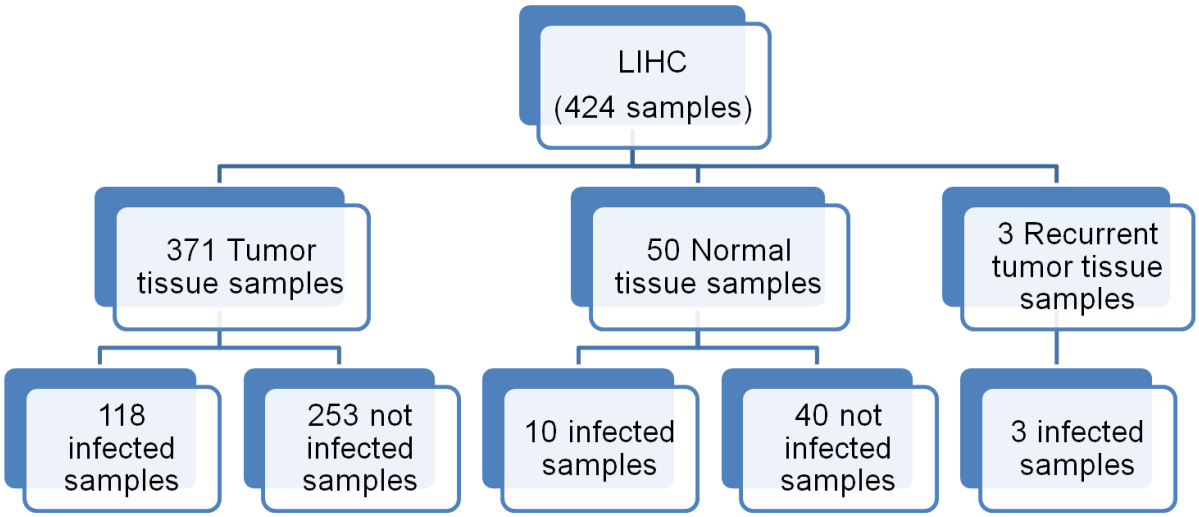


Figure 9. *RNAseq samples distribution through tissue type and infection results in samples from LIHC cancer.*

There were 50 NT samples to compare with the TP, and of these, two non-infected in NT were infected in TP, while three infected in NT were non-infected in TP. All recurrent tumor tissue samples were infected. The results obtained here for the viral infection in LIHC (Figure 10) meet the expected rates presented by Tang et al. in 2013, when analyzing a smaller subset of samples. Still, a low proportion of 8.89% of the samples, although considered as non-infected, have viruses detected in ppm between 2 and 10.

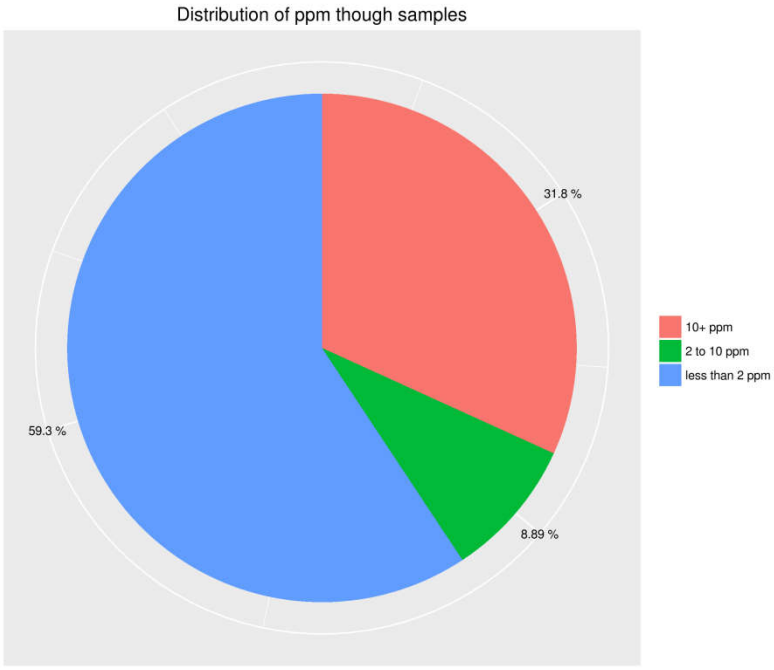


Figure 10. *Percentage distribution of viral infection in LIHC samples. Each sample was represented by the virus with the higher ppm and only tumour tissue samples were considered in this graphic. Virus with ppm lower than 10 were not considered as infecting a sample.*

In Figure 11 it is possible to see in more detail the distribution of the main viruses and their ppm in patients that have at least one virus attaining a ppm>2. Almost all of the 31.8% infected samples are infected by HBV (usually with high ppm), and less than 10% of those have HCV or HHV as the main infecting viruses (see also Table 3 in annexes). Adeno-associated virus (HAdV) was also observed in six samples, and in two of them, strain HAdV-2 attains a frequency between 100-300 ppm. Seven strains of HHV were observed in this sample, most with ppm inferior to 10, so we could not consider these viruses as infecting the samples, and only one had HHV5 strain Merlin with ppm between 10 and 50, and two other had HHV4 with ppm between 100 and 300. HHV1 strain 17 was the most frequent across patients, although always in small ppm. Three strains of HCV were observed, which genotype 1 attained infection rates in four samples (three with ppm between 50 and 100, and one between 100 and 300).

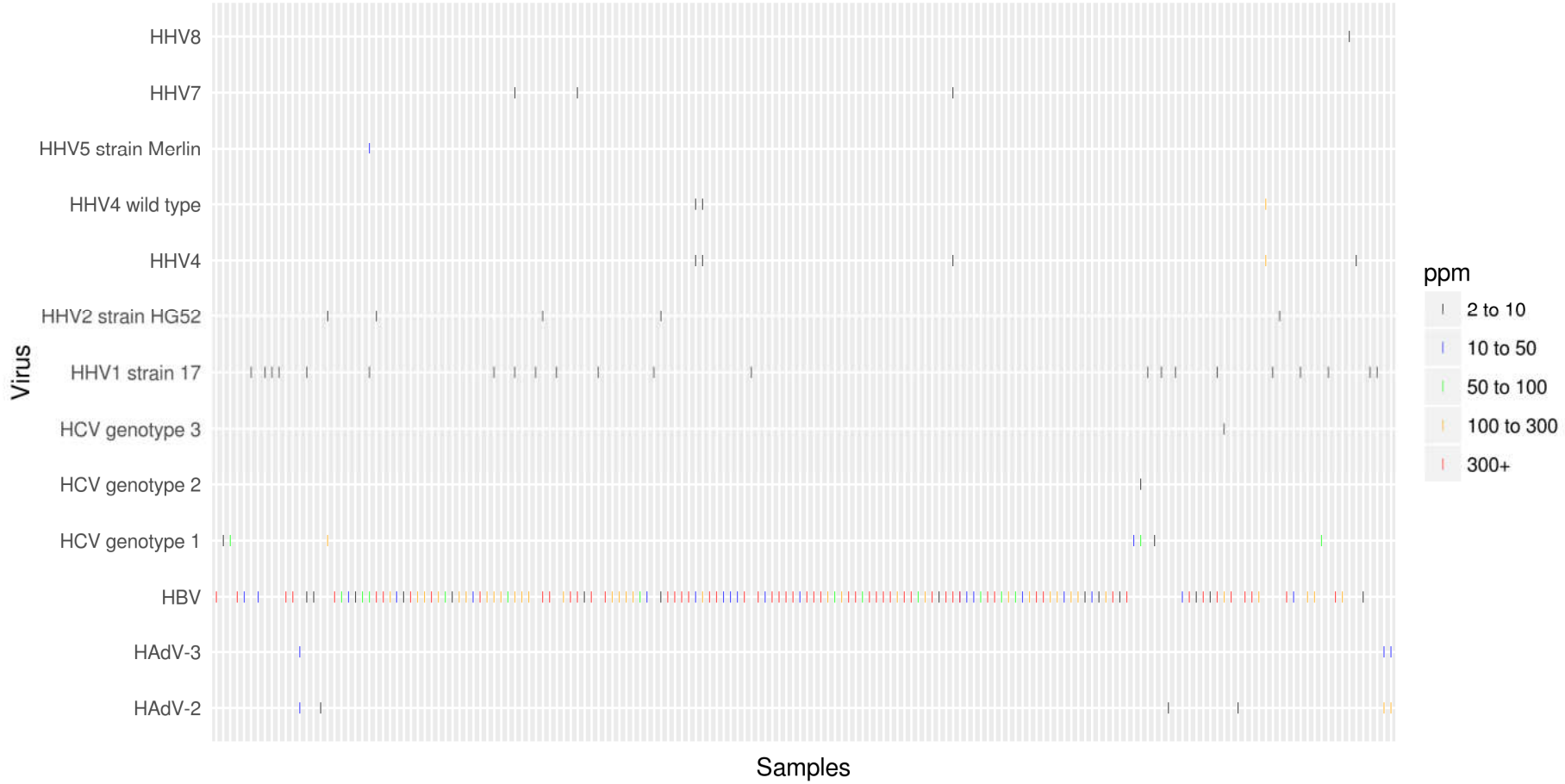


Figure 11. *Viral presence distribution in LIHC cancer samples by ppm. Each column represents one sample. Viruses with ppm smaller than 2 were not considered.*

The PCA test to evaluate clustering between TP and NT (Figure 12A) indicate a clear separation between the two groups, supporting a big difference in expression pattern in tumorigenesis. The PCA test between infected or non-infected groups (Figure 12B) does not show any differentiating clustering between the two categories, with only a slight tendency for the infected samples to spread in the upper side of the PCA.

A B

Figure 12. *PCA of LIHC samples. (A) Comparing Samples from tumor and normal tissue. (B) Comparing Samples from infected and not infected samples from TP samples only.*

Despite the not clear distinction in expression profiles between infected and non-infected groups in the PCA, we performed a GSEA analysis, as integrating expression of genes in metabolic pathways can be more informative: in PCA analysis, significant differences in a low number of genes may not be enough in the total of 20,000 genes, but if these alterations are in the few genes that make a metabolic pathway, they can reach statistical significance. The results presented in Figure 13 show that the enriched pathways in the infected group are related with cell division, DNA and RNA replication and repair, in all the reference lists of metabolic pathways. No pathway directly related to immune response is enriched in the infected group in LIHC. It seems that the common feature in cancer cells of constant replication and constant need of energy is significantly increased in the infected group when compared with the non-infected one.

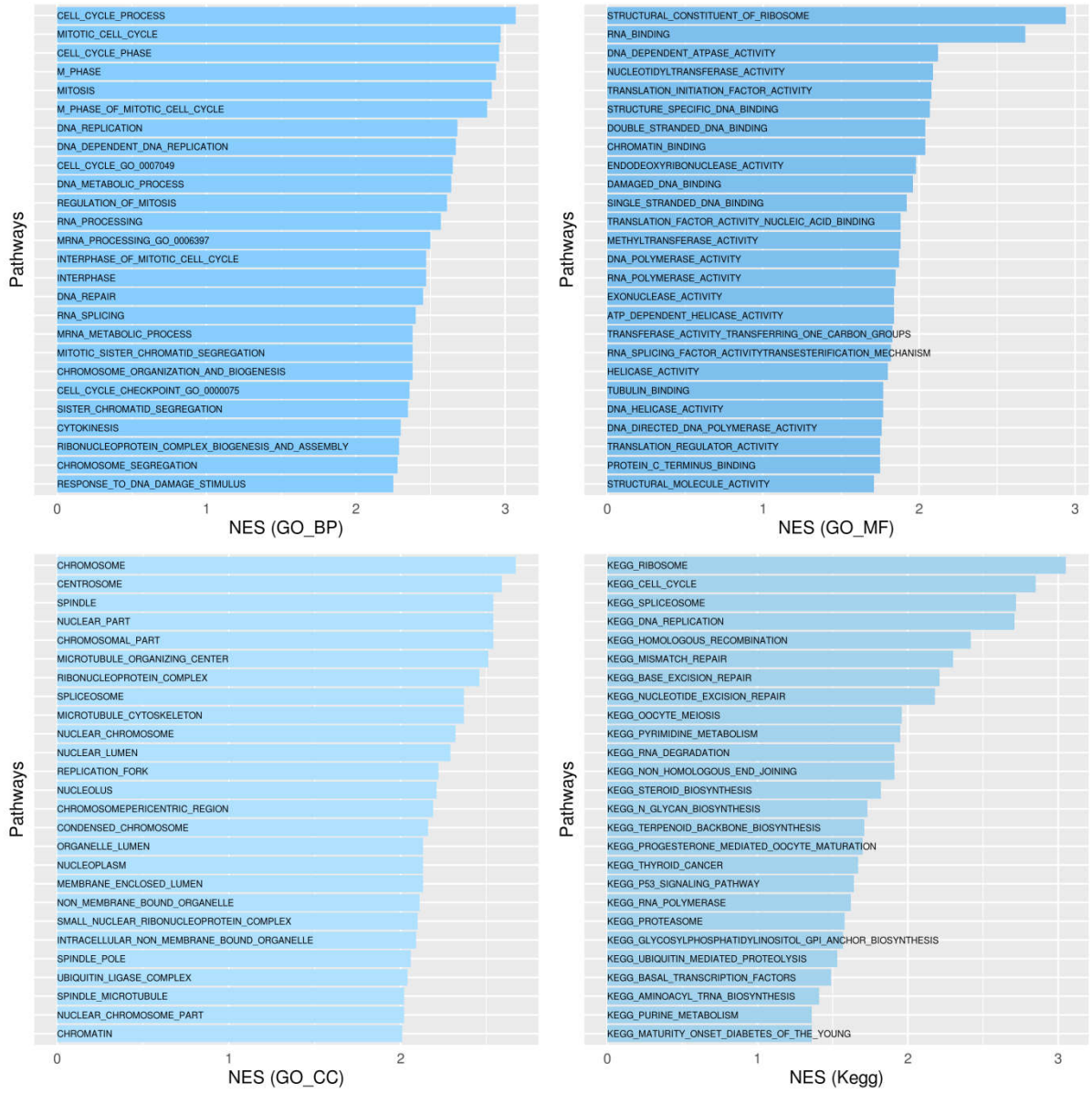


***Figure 13**. GSEA results for LIHC when comparing infected and non-infected samples and using the Gene Ontology Biological Process (GO_BP), Molecular Function (GO_MF) and Cellular Component (GO_CC), and the KEGG references lists. For each graphic 26 metabolic pathways with FDR smaller than 0.25 were picked and then sorted by NES value. The positive NES values mean that these are more expressed in the infected than in the non-infected.*

For LIHC, the more even number of samples in the infected and non-infected groups allowed us to evaluate the effect of infection in the amount of somatic mutations. The curated file downloaded from TCGA database contained information for 193 samples where 45 of them were infected and the others 148 samples were non-infected. The number of somatic mutations per sample varied extensively, from 40 to 1800 mutations per sample, corresponding to a total of 53,000 somatic mutations found in 193 LIHC samples. The total count and the ratios of different classes of somatic mutations in the two groups are displayed in Figure 14 and Table 4 in annexes.

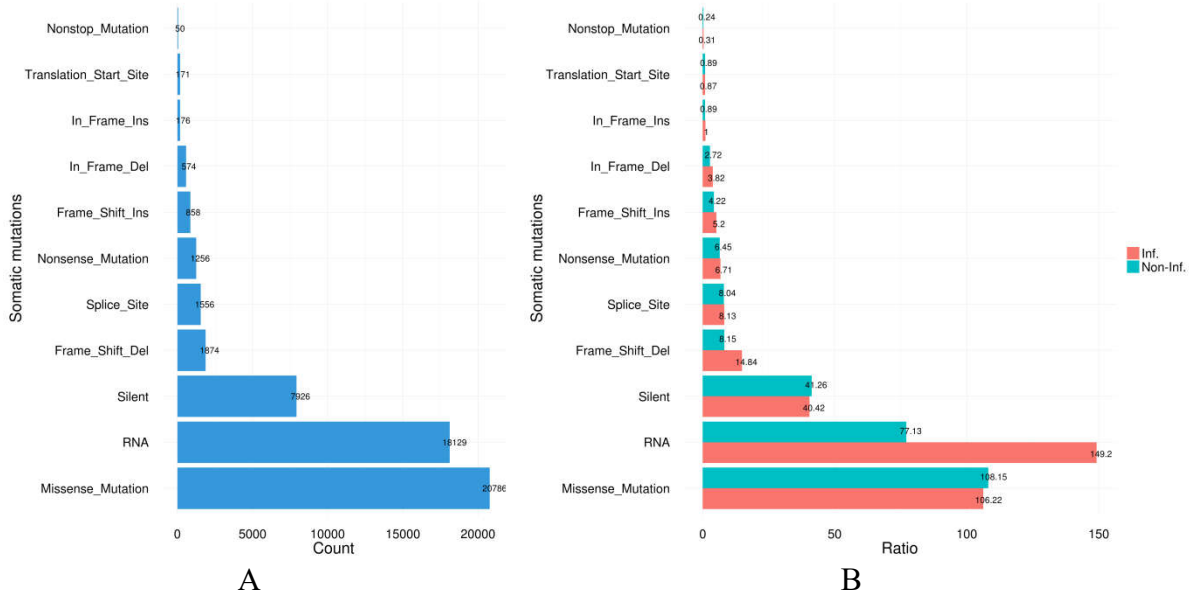


Figure 14. *LIHC Somatic Mutations. (A) Global count of somatic mutations in 193 LIHC samples. (B) Each somatic mutation type ratio (number of mutations found and divided by the number of sample) per infected and non-infected samples.*

In the 193 samples, 11 types of somatic mutations were found, with silent, RNA and missense being the most common ones. In opposition, in frame insertion, translation start site and nonstop mutations where rare. Globally, all mutations have similar ratio values in both groups, indicating that they are equally common in both infected and non-infected samples, although the tendency is for a slightly higher value in infected patients. The big exceptions are RNA and frame shift deletions mutations, which have a double ratio in infected compared with non-infected samples (statistically significant for the RNA class; two-tailed Fisher exact test p=0.017).

Then we performed a G:Cocoa analysis in the mutations occurring in coding genes, comparing the lists of ranked hit genes between the infected and non-infected groups. G:Cocoa results for LIHC (Figure 15) show that infected samples had a bigger number of significantly somatic mutated-hit pathways than non-infected samples. Most of these pathways are related with cancer and cell signalling, structural proteins on membranes, cell junction, DNA and RNA replication, transcriptional and energy mechanisms. These processes are known to be affected in cancer cells, which have a high rate of cell division. But more interesting, a few of the significant pathways in the infected group are related with immune response: phosphatidylinositol phospholipase C activity; inflammatory mediator regulation of TRP channels; leukocyte aggregation; somatic diversification of immune receptors. Maybe somatic mutations in the genes involved in these pathways render these patients more susceptible to the infection. The hit-mutated genes in the immune-related pathways are represented in Table 5 in annexes.

Figure 15. *Pathways having a significant amount of genes hit by somatic mutations in infected and non-infected LIHC groups obtained through G:Cocoa when running against Gene Ontology Biological Process (BP), Molecular Function (MF) and Cellular Component (CC), and the KEGG references lists.*

4.3. Head and neck squamous cell carcinoma (HNSC)

For the 566 HNSC RNAseq samples available at TCGA database, 520 were TP being 86 infected and 434 non-infected (16.5% infection rate), 44 NT with 3 infected and 41 non-infected, and 2 metastic tissue samples one infected and the other non-infected (Figure 16; Table 6 in annexes). Most of the samples having both tumor and normal tissue had concordant infection status except for four samples (two infected and two non-infected TP samples had its analogous NT samples with opposite viral infection presence).

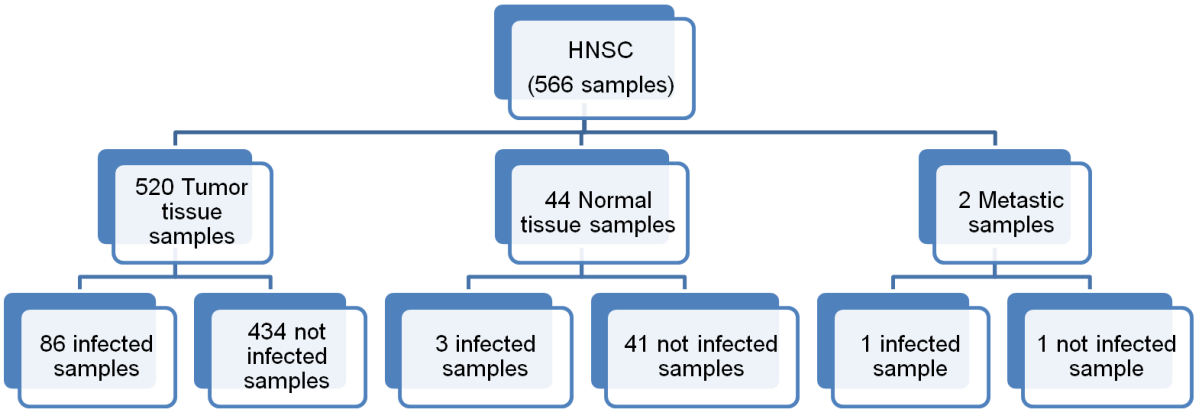


Figure 16. *RNAseq samples distribution through tissue type and infection results in samples from HNSC cancer.*

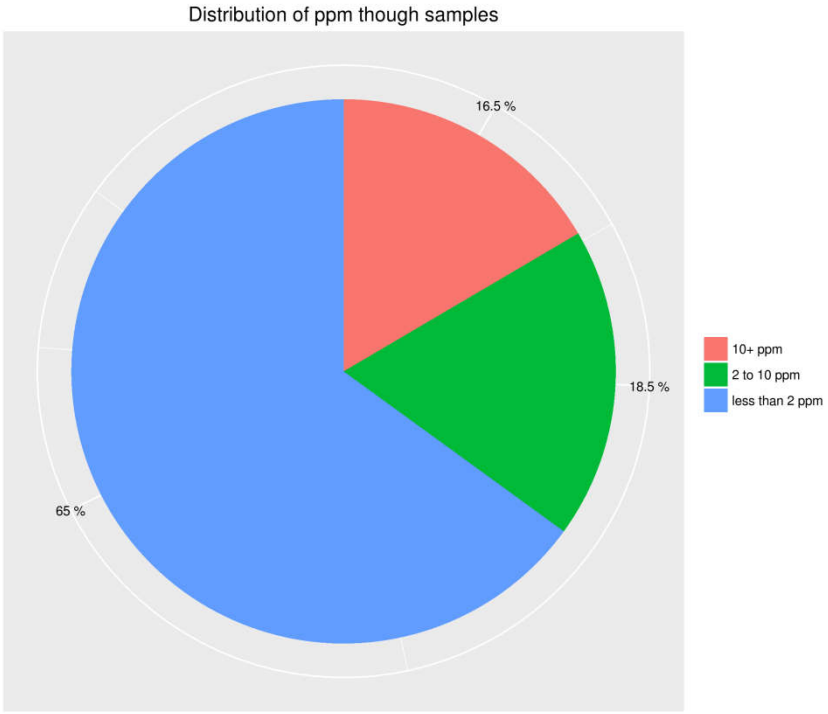


Figure 17. *Percentage distribution of viral infection in HNSC samples. Each sample was represented by the virus with the higher ppm and only tumour tissue samples were considered in this graphic. Virus with ppm lower than 10 were not considered as infecting a sample.*

The 16.5% infection rate in HNSC, the smallest in the three types of cancer studied in this work, is concordant with values estimated previously in a very limited subset of this cohort (Tang et al. 2013).

From all viruses found, HPV16 is the most frequent (80% of infected samples) and where the highest values of ppm are attained, with a big proportion of patients having ppm values above 300 and between 100 and 300. The $HPV16^{+}$ cases were mainly observed in tonsil (48.5%) and base of tongue (19.1%), corresponding to 73.3% and 48.1% of the cases in these tissues. Another HPV strain infected some patients, HPV26, but ppm values were in general lower, around 10 to 50 and 50 to 100. Two other HPV, HPV53 and HPV32, were observed only in one sample each (Table 6 in annexes).

A total of seven HHV strains were observed, with low ppm values that would not be considered as infecting the patient, in most of the cases. HHV1 strain 17 was the most common among all samples, but there are only three samples where this virus had ppm above 10.

Apart from HHV and HPV viruses, there were also other two viruses present (HBV and HAdV-C). The HBV virus was only present in two samples and even if it was considered as infecting those samples, its ppm were not much higher than 10. In the other hand, HAdV-C virus were present in more samples but none where this virus had ppm above 10.

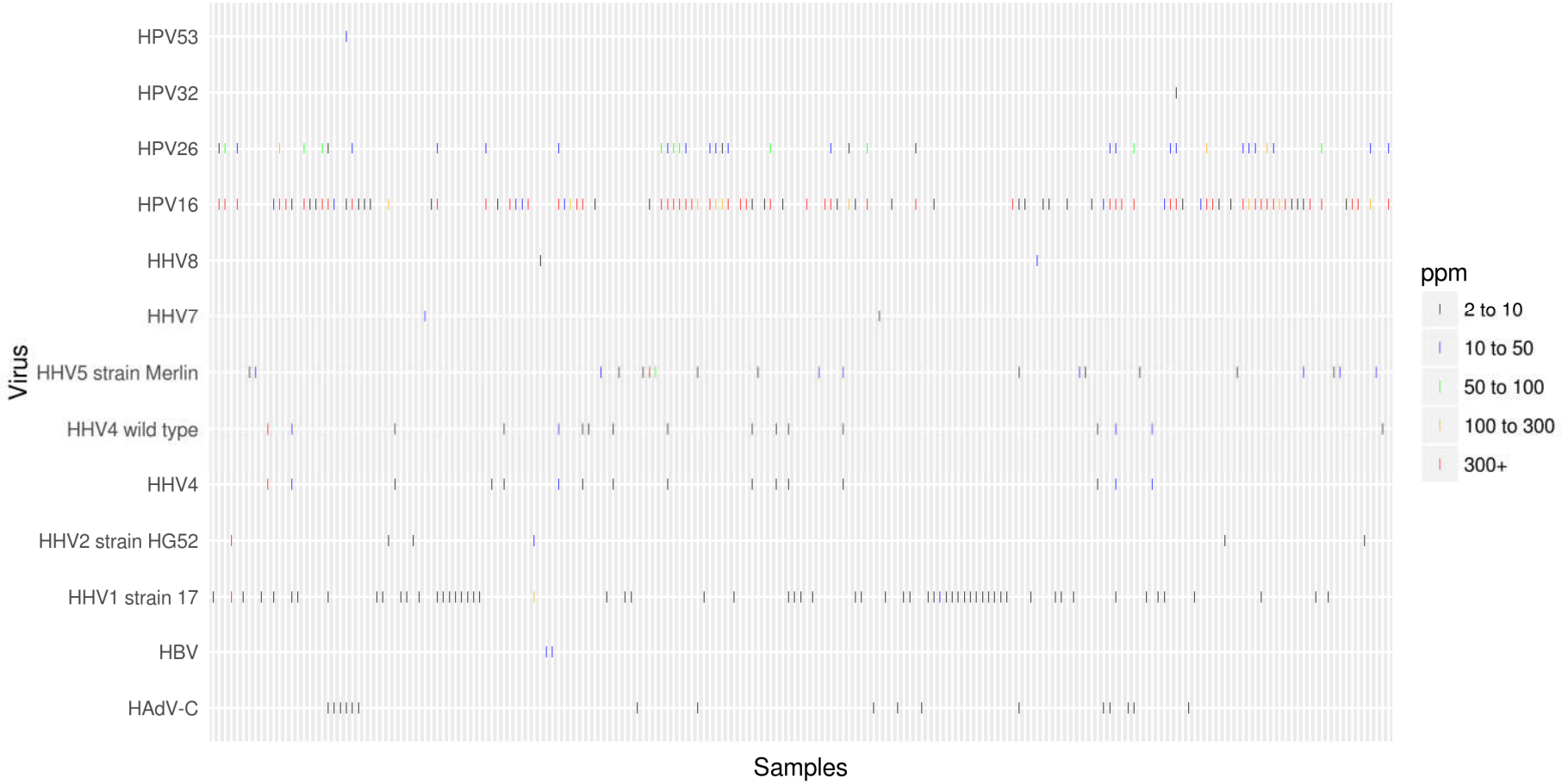


Figure 18. *Viral presence distribution in HNSC cancer samples by ppm. Each column represents one sample. Viruses with ppm smaller than 2 were not considered.*

The PCA on the comparison between TP and NT samples (Figure 19A), did not reveal clear clustering, with TP samples being dispersed and NT samples spreading among them and slightly displacing to the left side of the plot. This seems to indicate slight differences in gene expression profile between TP and NT samples. The same is true for the PCA comparison between infected and non-infected samples (Figure 19B), with samples spreading all over the plot, and just a slight higher frequency of infected samples in the lower left corner of the PCA. Again, this seems to indicate the existence of a few differences in gene expression profile between infected and non-infected samples

A B

Figure 19. *PCA of HNSC samples. (A) Comparing Samples from tumor and normal tissue. (B) Comparing Samples from infected and not infected samples from TP samples only.*

The main genes responsible for the variability in PC1, in decreasing order, were: *KRT19*, *KLK5*, *KRTDAP*, *KRT1*, *DEFB103B*, *NTS*, *CDSN*, *SPRR2G*, *DSG1*, *WNK2*, *WFDC12*, *FAM25A*, *LCE3D*, *MYO3A*, *LCE3E*, *PNLIPRP3*, *S100A7*, *SPINK6*, *PCDH19*, *KRT14*, *FAM3B*, *CRCT1*, *NTRK2*, *SPRR2B* and *S100A7A*.

The GSEA results (Figure 20; genes in main pathways are listed in Table 7 in annexes), for the GO_BP reference list, indicated that pathways differentially more expressed in infected than in non-infected groups are related to DNA replication and organization, and cell division, with none related to viral infection or to lipid metabolisms. Similar results were found when using GO_MF, but it included lipid kinases activity pathway. When using GO_CC, the nucleus and chromosome are the main cellular components identified, but also an immunological synapse is significant. Regarding the results for KEGG reference list, in concordance with the previous lists, some pathways related to cell division and DNA repair are observed, but the

signal for differential expression in immune-related pathways is stronger (T- and B-cell receptor signalling, primary immunodeficiency, intestinal immune network for IGA production).

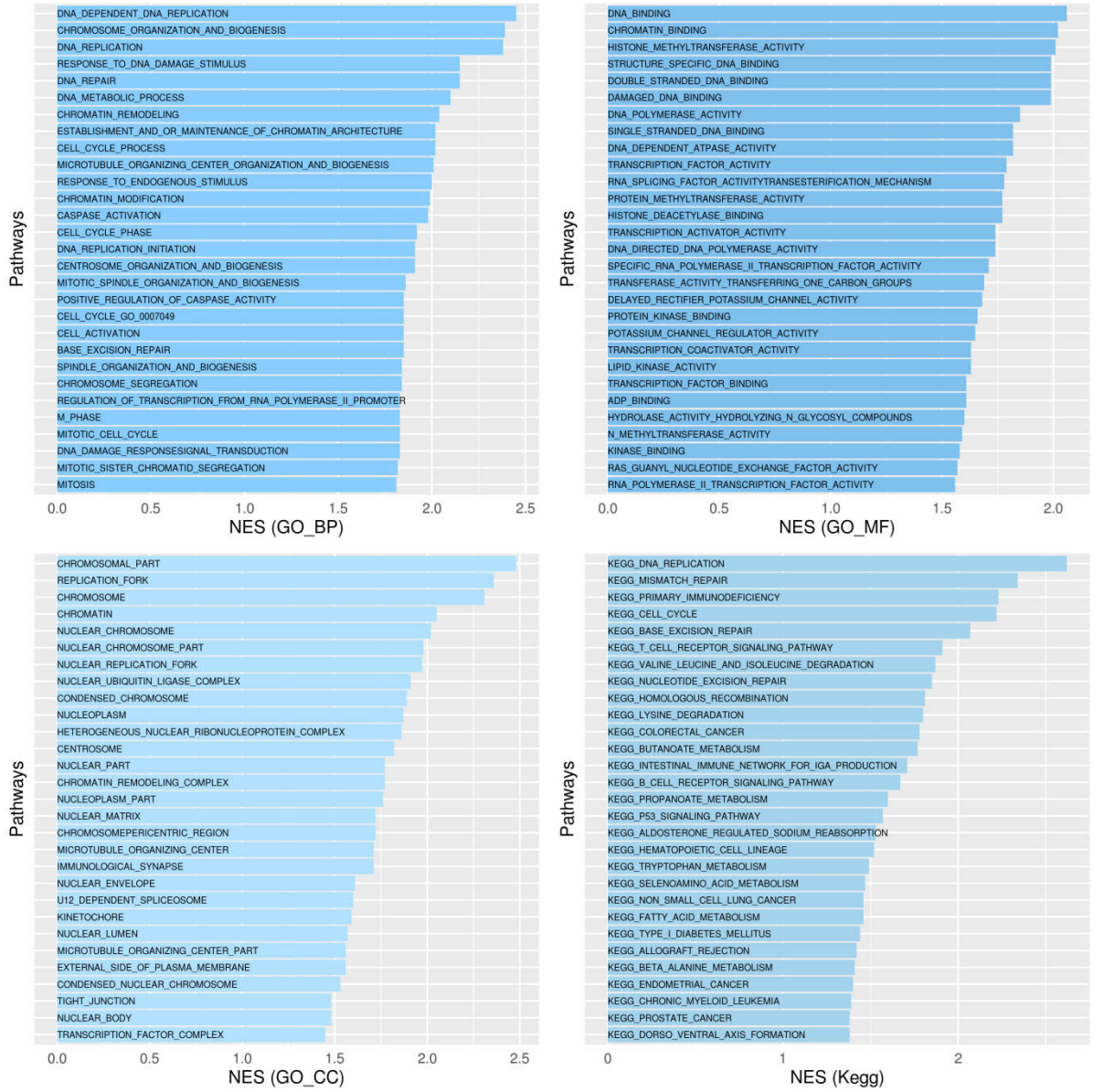


Figure 20. *GSEA results for HNSC when comparing g infected and non-infected samples and using the Gene Ontology Biological Process (GO_BP), Molecular Function (GO_MF) and Cellular Component (GO_CC), and the KEGG references lists. For each graphic 29 metabolic pathways with FDR smaller than 0.25 were picked and then sorted by NES value. The positive NES values mean that these are more expressed in the infected than in the non-infected.*

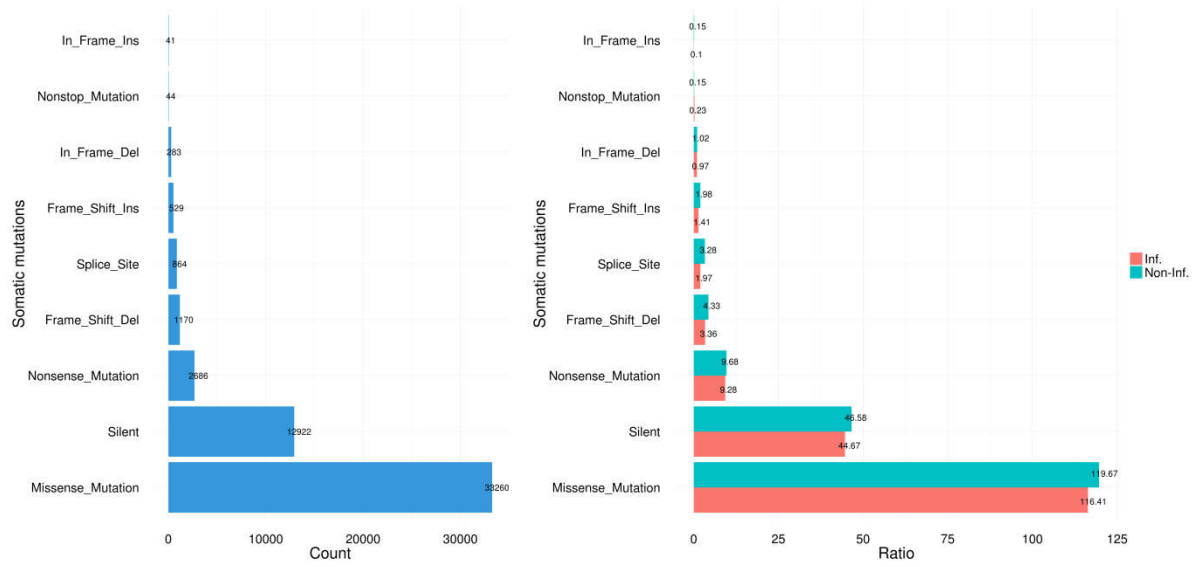


Figure 21. *HNSC Somatic Mutations. (A) Global count of somatic mutations in 279 HNSC samples. (B) Each somatic mutation Ratio (number of mutations found and divide them by the number of sample) per infected and non-infected samples.*

For HNSC, TCGA website provides information for somatic mutations in less than half of the initial samples, amounting to 279 samples where 39 were infected and 240 were non-infected samples. It was possible to identify (Figure 21 and table 9 in annexes) that silent and missense mutations were the most frequent mutations in HNSC cancer appearing over 12900 and 33200 times, respectively, in all 279 samples. On the other hand, in frame insertions and nonstop mutations were very rare in these samples appearing only around 40 times each. Regarding the ratio of those mutations among infected and non-infected samples, they are identical. The G:Cocoa analysis of the somatic mutations occurring in coding genes (Figure 22) indicated that in both infected and non-infected groups many similar pathways directly related to cell maintenance and division mechanisms were found, in accordance with higher rate of cell replication in tumor tissues. Regarding the results in the infected group, some pathways are related to viral infection, as lymphocyte activation and Fc gamma R-mediated phagocytosis. Lymphocytes act as principal response to viral infection in an organism and phagocytosis plays an essential role in host-defense mechanisms through the uptake and destruction of infectious pathogens. The hit-mutated genes in the immune-related pathways are represented in Table 8 in annexes.

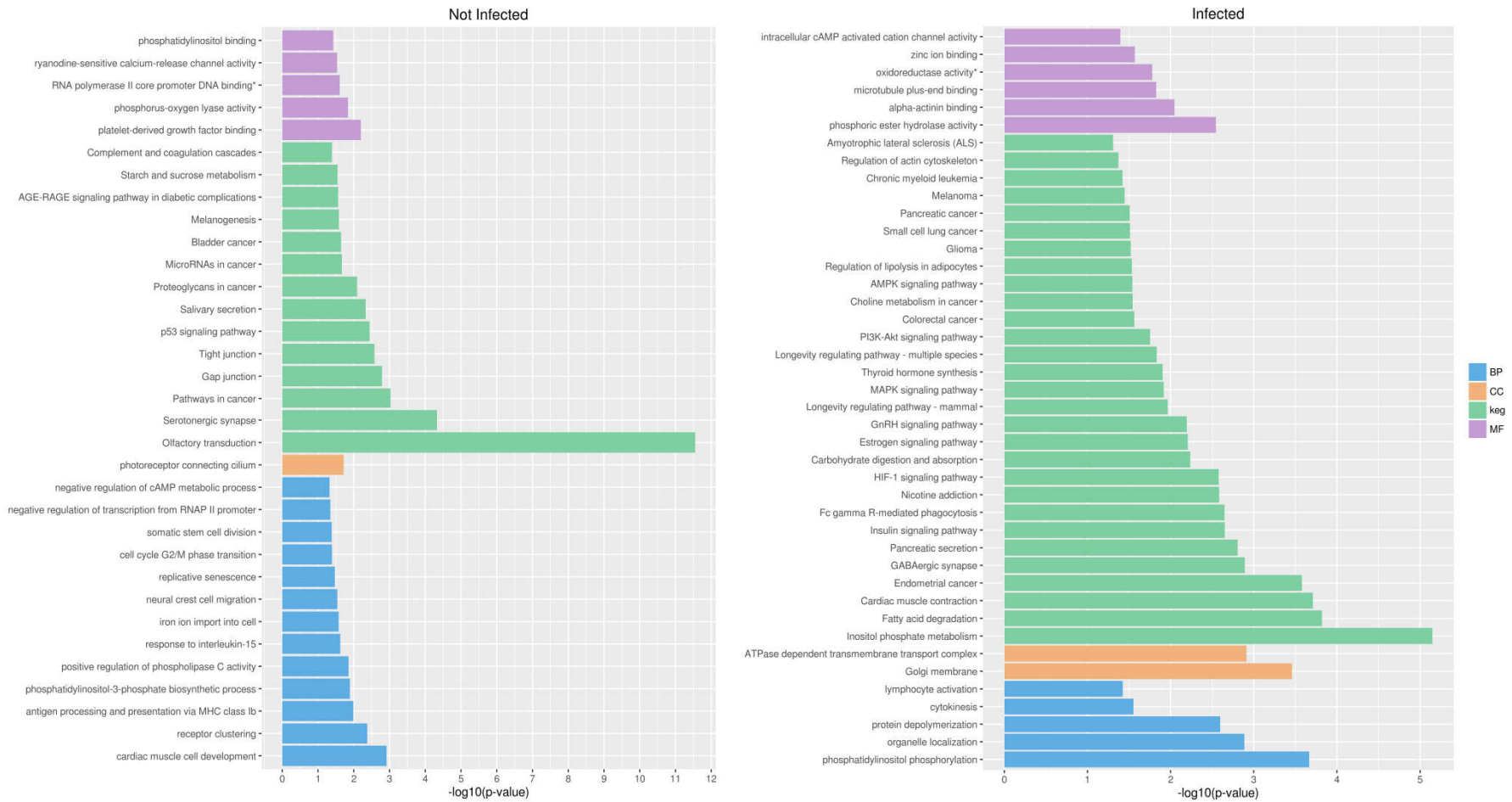


Figure 22. *Pathways affected by somatic mutations in infected and non-infected HNSC samples obtained through G:Cocoa when running against Gene Ontology Biological Process (BP), Molecular Function (MF) and Cellular Component (CC), and the KEGG references lists.*

Furthermore, a single whole genome sample was downloaded in order to confirm and detect viral insertion in the host genome, by using the VirusSeq tool. The sample was from a patient where the HPV16 virus was highly expressed, and which Tang et al. (2013) confirmed has having the virus inserted in the host genome. The whole genome had approximately 200 GB, and this this tool used around 50GB RAM memory, which with the server we had available took around two weeks to finish.

Table 3. *VirusSeq results of HNSC sample TCGA-BA-4077-01B. Confirmation of HPV16 insertion on an infected sample and its correspondent expressed genes and location on the host genome.*

Viral Transcript	Host Gene	Discordant Read Pairs	Virus Chr. ID	Virus Integration Location	Gene Chr. ID	Gene Integration Location
HpV16gp2_E7	RAD51B//intron7	331	25	581007	14	68699686
HpV16gp4_E2	RAD51B//intron7	46	25	583997	14	68683564
HpV16gp4_E2	RAD51B//intron7	103	25	584022	14	68741478
HpV16gp5_E4	RAD51B//intron7	440	25	583794	14	68741301
HpV16gp6_E5	RAD51B//intron7	10	25	584257	14	68683629

We confirmed that the HPV16 virus was inserted in the genome of the patient. That the viral genes inserted were E2, E4, E5 and E7, which were inserted on the chromosome 14, in the intron 7 of *RAD51B* gene of the human genome (Table 3). These results replicated the ones obtained by Tang et al. (2013), but due to limitations of server, we could not pursue in checking viral insertion in all the infected samples we characterised here for the three tumor types.

4.4. Direct comparison between all cancer results

The observation that, in CESC and HNSC, a big portion of samples are infected by HPV enables to test if the same type of virus can influence significantly on the host gene expression profile, independently on the infected tissue. We thus performed a PCA analysis for CESC and HNSC infected samples and identifying the presence or absence of HPV (Figure 23).

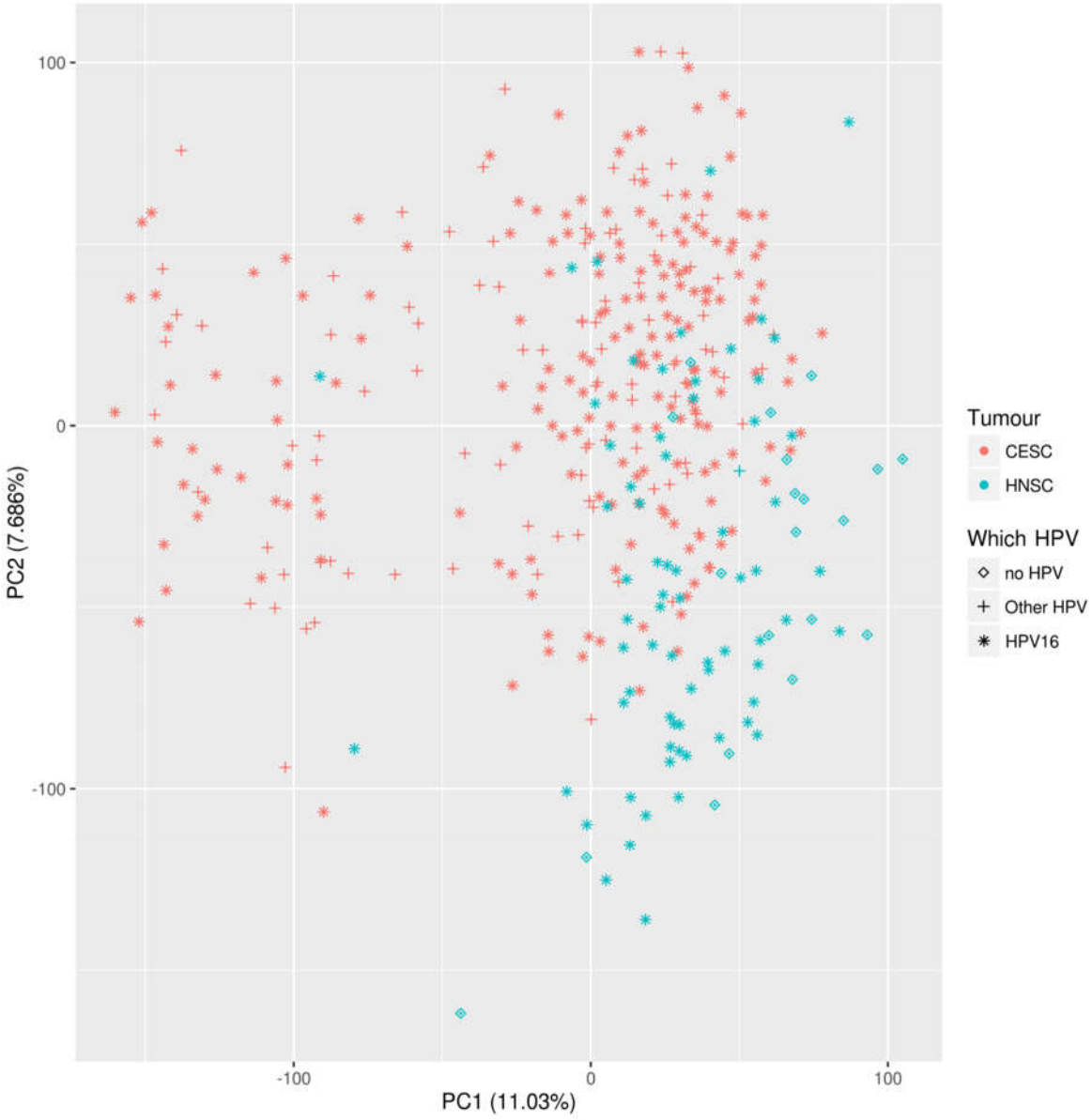


Figure 23. *Comparison between CESC and HNSC cancer expression profiles. Only infected samples from each cancer were used and were identified by the presence or absence of any HPV virus.*

The two cancers tend to form clusters, but there is still a considerable overlap between CESC and HNSC infected HPV samples along the right side of the PCA. Joining these results with the ones observed in Figure 7B this side of the plot is mainly formed by CESC samples infected

by HPV16, which is the dominant HPV strain observed in HNSC infected samples. It seems thus that this strain of HPV determines a response of the host that involves a distinct host gene profile that prevails across cervix and head and neck tissues.

The main genes responsible for the variability in PC1, in decreasing order, were: *KRT14*, *DSG3*, *SPRR1B*, *SPRR2A*, *IVL*, *SPRR1A*, *DSC3*, *CLCA2*, *SBSN*, *SPRR2D*, *TMPRSS11D*, *CALML3*, *SPRR2E*, *KRT6C*, *RHCG*, *SERPINB13*, *KRT6A*, *LASS3*, *PKP1*, *SPRR3*, *KRT6B*, *BNC1*, *GBP6*, *MUC5B* and *TP63*.

In these cancers, we selected five samples from each type which were HPV16^{+}, in order to check which viral genes are more expressed by using HTSeq tool.

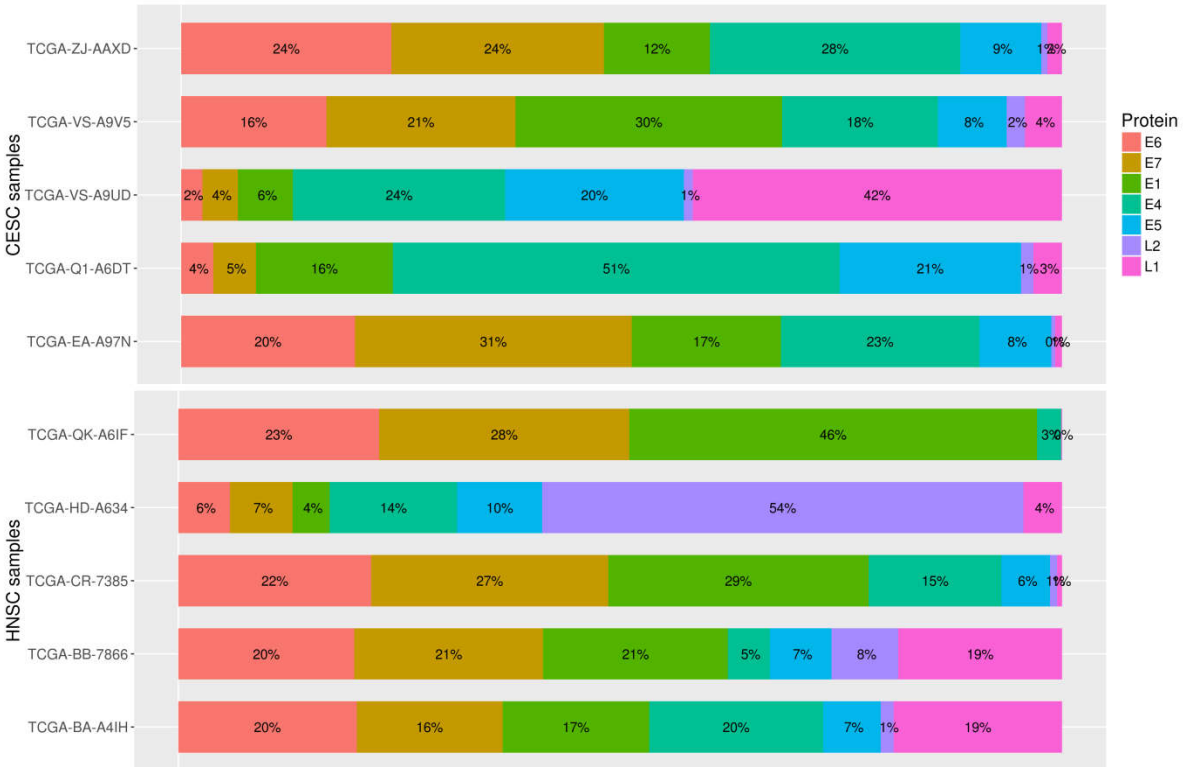


Figure 24. *Distribution of HPV16 expressed genes when infecting CESC and HNSC samples.*

The HPV16 genome contains 10 protein coding genes (CDS), whose functions are presented in Table 10 in annexes. From the results obtained through HTSeq (Figure 24; Table 11 in annexes) it is possible to say that the two genes coding the transforming proteins E6 and E7 are highly expressed in almost all samples, in both tissues. E1 and E2, which play an important role in viral DNA replication are differentially expressed, with E1 being expressed in all samples and E2 non-expressed. The transforming protein E5, which contributes to cell transformation and polyploidy cells through endoreplication, is more expressed in CESC than in HNSC. While the genes for the minor capsid protein L2 and major capsid protein L1, playing

an essential role in virus assembly by recruiting viral components, have usually low expression except in two samples, L1 being dominantly expressed in one CESC sample and L2 dominating in one HNSC sample.

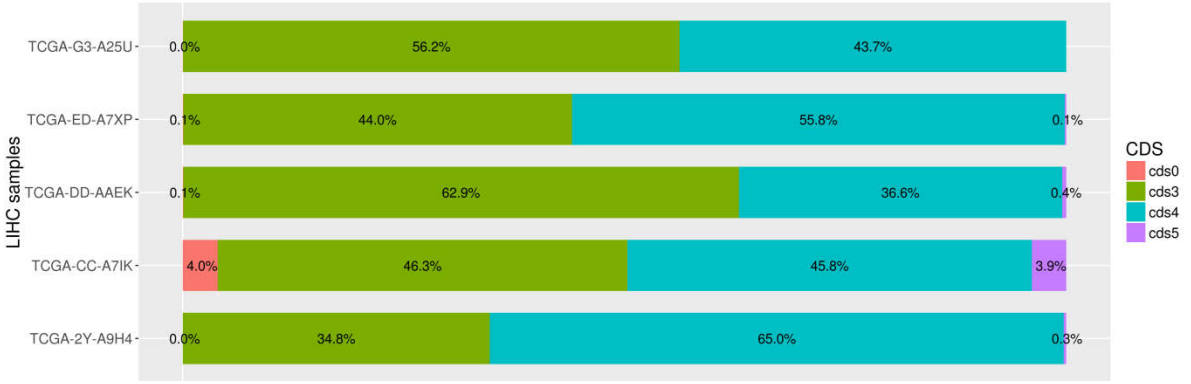


Figure 25. *Distribution of HBV expressed genes when infecting LIHC samples.*

We also performed a similar analysis in five LIHC samples HBV^{+}. HBV only has seven CDS (Table 12 in annexes). In this cancer the variability of expressed CDS is far smaller (Figure 25; Table 13 in annexes). All samples have a big percentage of CDS3 and CDS4 expression, which code the small envelope protein and the X protein, respectively. Both are very important for viral infection especially the X protein, since it promotes cell cycle progression and inhibits tumour suppressor protein. Other CDS like CDS0, which codes polymerase protein, and CDS5, responsible for assembly of pre-capsid proteins, are expressed in very small percentages. Finally, the remaining CDS (CDS1, CDS2 and CDS6) coding envelope protein assembles were completely undetectable.

5. CONCLUSION

The availability of extensive RNAseq datasets in TCGA for CESC, LIHC and HNSC allowed us to investigate the infection rates in these tumors, in sample sizes between 309 and 566. The infection rates we obtained accorded with values described previously in more limited datasets (Tang et al. 2013): 94% for CESC; 32% for LIHC; and 17% for HNSC. At the same time, we were able to characterise in great detail the viral diversity observed in each human sample, even for low present viruses. Most infected CESC samples bear several HPV strains, although HPV16 is by far the most common in each sample and overall in the dataset. LIHC infected samples are homogeneously infected by HBV, while most HNSC infected samples are mainly infected by HPV16. HHV1 strain 17 is common in all three tissues, especially so in HNSC, but usually does not attain infection levels.

When investigating the impact of infection upon the host gene expression, a clear sign of enrichment in immune-related pathways was observed in the infection group in CESC, less so in HNSC, and not at all in LIHC. The signal in CESC seems to indicate an active immune response against the virus, as the pathways detected as enhanced were: cellular defence response, defence response, immune response, response to virus, inflammatory response, immune system process, endosome transport, JAK-STAT cascade, response to other organism, T-cell activation, pathways involving interleukins, cytokines and chemokines, proteasome regulation, natural killer cell mediated cytotoxicity, primary immunodeficiency and antigen processing and preservation. These results contradict claims that cervical cancer arises rarely as a result of infection, with most infections being cleared by the host without clinical symptoms, as in general, HPV infections evade both adaptive and innate immune responses, with the life cycle being totally intra-epithelial, without viraemia, cell lysis or inflammation (Egawa et al. 2015). Till this work, no proper GSEA test had been applied to the comparison between infected and non-infected groups, given the extremely high infection rate not enabling enough number of non-infected patients for proper statistical evaluation (Adams et al. 2014). The larger dataset analysed here, allowed us to overcome this limitation, and to provide evidence on the contrary. We observed that those immune-related pathways incorporate genes such as *MICB*, *HLA-B*/*G*, *CCL5*, *CCR5*/*6*/*7*, *KLRC2*/*3*/*4*, *IL1RL2*/*IL10RB*/*IL12B* (Table 2 in annexes) These genes play a leading role in the immune response to viral infection. In HNSC, the few immune related enriched pathways in the infected group were limited to: immunological synapse, T- and B-cell receptor signalling, primary immunodeficiency and intestinal immune network for IGA production. These pathways involved mostly major histocompatibility complex (*HLA-*

DMA/DMB/DOA/DOB/DPB1/DQA2/DRA/DPA1/DQA1), tumor necrosis factor receptor superfamily (*TNFRSF13B/13C/17*), PIK3 (*PIK3R1/R3/R5/CA/CB/CG*) and mitogen-activated protein kinase (*MAPK13/MAP2K7/3K8/3K14*) genes, which also play a role in signalling control of cell cycle (Table 7 in annexes). Recently, Yan et al. (2016) by analysing HNSC TCGA mRNA and miRNA datasets, and performing differential analyses between tumor and normal tissues, observed several immune-related pathways as overexpressed in the tumor tissue (as defense response, response to virus, and primary immunodeficiency) (Yan et al. 2016). These results of a clear signal of immune-response involvement in CESC need to be confirmed in an independent work.

Another category of pathways enriched in CESC infected group was more compatible with integration of viral genome in the host genome: pathways related with chromosomes, DNA replication, mismatch repair, cytosolic DNA sensing pathway. Genes of the family minichromosome maintenance complex, as *MCM2/4/5/6*, which are highly conserved proteins involved in the initiation of eukaryotic genome replication, polymerase DNA catalytic subunits (*POLA1/2*, *POLD1/2/3/4*, *POLE2/3/4*) also involved in the initiation of DNA replication, are amongst the significantly enriched expressed genes. These pathways are also the dominantly enriched in the infected group in LIHC and HNSC, through the same genes and also other families as the cyclin-dependent kinase inhibitors (*CDKN2D*), that act as negative regulators of proliferation of normal cells and, in the case of *CDKN2A*, a well-known biomarker for HNSC infection (Zhang et al. 2016). This seems to indicate that integration in the host genome is very important in the three cancer types. Tang et al. (2013) showed that the integration of the viral genome in the host genome is very high for HPV18 (100%), HPV16 (59%) and HBV (77%), which are the dominant viruses in infected CESC, LIHC and HNSC samples. It is known that for HPV16, the main trigger for the transformation of the epithelial cells is the overexpression of the viral genes E6 and E7, usually integrated in the host genome, which lose control from the viral E2 protein that is mostly not integrated or disrupted (Adams et al. 2014; Egawa et al. 2015). E6 and E7 proteins bind, respectively, to and inactivate the tumor suppressor proteins TP53 and RB1, and thus host cells avoid apoptosis and grow in an uncontrolled manner (Zhang et al. 2016). Our analyses of the expression of HPV genes confirmed the overexpression of E6 and E7 genes and almost no-expression of E2, both in the samples from CESC and HNSC. However, in the only integration confirmation we performed, we saw that E2 was also integrated in the host genome in this case. We also detected samples were instead of E6

and E7 overexpression, there was overexpression of L1 or L2 genes, which may indicate that these samples have a more active infection, with need of production of capsid for new virions.

Still in CESC, one top enriched pathway in the infected group was keratinocyte differentiation, which perfectly matches the fact that HPV infection and replication occurs, respectively, in the proliferating and differentiated keratinocytes of the epithelium (Adams et al. 2014). Genes for small proline rich proteins (*SPRR1A/B*) and involucrin (*IVL*), which are cross-linked envelope proteins of keratinocytes, as well as transglutaminase (coded by *TGM1*) which cross-links those proteins to membrane proteins, are overexpressed in that pathway. Interestingly, these genes are amongst the ones responsible for splitting the two clusters in PCA comparing infected and non-infected groups in CESC, which we confirmed that is also associated with a separation (although not perfect) between $HPV16^-$ and $HPV16^+$ samples. In fact, the main genes responsible for PC1 in that figure (Figure 7B) are the ones coding for the already mentioned small proline rich proteins (*SPRR1A/B* and *SPRR2A/D*) and involucrin (*IVL*), as well as keratins (*KRT5/6A/6B/6C/14*), *SERPINB13* (that may play a role in the proliferation or differentiation of keratinocytes), and other proteins related with cell-adhesion. These genes are identical to the ones governing PC1 in HNSC (Figure 19B) and CESC together with HNSC (Figure 23), showing how the tropism of HPV16 to keratinocytes (Egawa et al. 2015) is fundamental in driving the host gene expression in the two tissues. However, the signal of keratinocyte-related pathway is not detected in the GSEA analyses conducted in HNSC.

Given the viral integration in the host genome, which mostly occurs in known cancer genes such as *MYC*, *ERBB2*, *RAD51B* (as in the sample we analysed here), *MLL4* and the 13q22.1 region (Tang et al. 2013). This integration leads to increase in somatic mutations in the infected group, except for HNSC, given that tobacco is an important carcinogen in this tumor type and that TP53 mutations occur in HPV^- cases (Adams et al. 2014; Zhang et al. 2016). As there were few non-infected samples for comparison purposes in CESC, we did not conduct these tests in this tumor type. But we confirmed the higher amount of somatic mutations in the infected group, especially RNA and frame shift deletions, in LIHC, and the slight lower amount of mutations in HNSC infected group. In concordance, when performing enrichment pathway analyses, the LIHC infected samples had a bigger number of significantly somatic mutated-hit pathways than non-infected samples, while in HNSC, the number was similar in both groups. In LIHC, the pathways were mostly related with cancer and cell signalling but a few related with immune response, such as phosphatidylinositol phospholipase C activity (mostly phospholipase C genes mutated), inflammatory mediator regulation of TRP channels (many *MAPK*

genes), leukocyte aggregation (including *TNFSF* genes) and somatic diversification of immune receptors. In HNSC, in both infected and non-infected groups there were many similar pathways directly related to cell maintenance and division mechanisms, while in the infected group, some pathways are related to viral infection, as lymphocyte activation (including *MAPK* and *PIK3* genes) and Fc gamma R-mediated phagocytosis (*PIK3* genes). In accordance with our results, the comprehensive analyses performed by TCGA of the HNSC (Lawrence et al. 2015) showed that one dominant feature of the HPV^+ associated HNSC tumors is the presence of helical domain mutations of the oncogene *PIK3CA*. It seems that some somatic mutations in the infected LIHC and HNSC groups may render these patients more susceptible to the infection.

REFERENCES

Adams, Allie; Wise-Draper, Trisha; Wells, Susanne. 2014. "Human Papillomavirus Induced Transformation in Cervical and Head and Neck Cancers." *Cancers* 6 (3): 1793–820.

Anders, Simon; Pyl, Paul Theodor; Huber, Wolfgang. 2015. "HTSeq-a Python Framework to Work with High-Throughput Sequencing Data." *Bioinformatics* 31 (2): 166–69.

Broad Institute. 2016. "Picard Tools." *Http://broadinstitute.github.io/picard/ [Accessed 20 July 2016]*.

Canavan, Timothy P.; Doshi, Nipa R. 2000. "Cervical Cancer." *American Family Physician* 61 (5): 1369–76.

Chen, Chien-Jen; Hsu, Wan-Lun; Yang, Hwai-I; Lee, Mei-Hsuan; Chen, Hui-Chi; Chien, Yin-Chu; You, San-Lin. 2014. "Epidemiology of Virus Infection and Human Cancer." In *Viruses and Human Cancer*, edited by Mei Hwei Chang and Kuan-Teh Jeang, 193:11–32. Recent Results in Cancer Research. Berlin, Heidelberg: Springer Berlin Heidelberg.

Chen, Min-Shan; Li, Jin-Qing; Zheng, Yun; Guo, Rong-Ping; Liang, Hui-Hong; Zhang, Ya-Qi; Lin, Xiao-Jun; Lau, Wan Y. 2006. "A Prospective Randomized Trial Comparing Percutaneous Local Ablative Therapy and Partial Hepatectomy for Small Hepatocellular Carcinoma." *Annals of Surgery* 243 (3): 321–28.

Chen, Yunxin; Yao, Hui; Thompson, Erika J.; Tannir, Nizar M.; Weinstein, John N.; Su, Xiaoping. 2013. "VirusSeq: Software to Identify Viruses and Their Integration Sites Using next-Generation Sequencing of Human Cancer Tissue." *Bioinformatics* 29 (2): 266–67.

Consortium, T. G. O. 2001. "Creating the Gene Ontology Resource: Design and Implementation." *Genome Research* 11 (8): 1425–33.

Egawa, Nagayasu; Egawa, Kiyofumi; Griffin, Heather; Doorbar, John. 2015. "Human Papillomaviruses; Epithelial Tropisms, and the Development of Neoplasia." *Viruses* 7 (7): 3863–90.

Egloff, Ann Marie; Lee, Ju-Whei; Langer, Corey J; Quon, Harry; Vaezi, Alec; Grandis, Jennifer R; Seethala, Raja R; et al. 2014. "Phase II Study of Cetuximab in Combination with Cisplatin and Radiation in Unresectable, Locally Advanced Head and Neck Squamous Cell Carcinoma: Eastern Cooperative Oncology Group Trial E3303." *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research* 20 (19): 5041–51.

Heaton, Nicholas S.; Randall, Glenn. 2011. "Multifaceted Roles for Lipids in Viral Infection." *Trends in Microbiology* 19 (7): 368–75.

Hudson, Thomas J.; Anderson, Warwick; Aretz, Axel; Barker, Anna D; Bell, Cindy; Bernabé, Rosa R; Bhan, M K; et al. 2010. "International Network of Cancer Genome Projects." *Nature* 464 (7291): 993–98.

Irizarry, R. A.; Wang, Chi; Zhou, Yun; Speed, T. P. 2009. "Gene Set Enrichment Analysis Made Simple." *Statistical Methods in Medical Research* 18 (6): 565–75.

Jolliffe, Ian T.; Cadima, Jorge. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2065): 20150202.

Konan, Kouacou V; Sanchez-Felipe, Lorena. 2014. "Lipids and RNA Virus Replication." *Current Opinion in Virology* 9 (4): 45–52.

Langmead, Ben; Salzberg, Steven L. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Langmead, Ben; Trapnell, Cole; Pop, Mihai; Salzberg, Steven L. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3):

R25.1-R25.10.

Lawrence, Michael S.; Sougnez, Carrie; Lichtenstein, Lee; Cibulskis, Kristian; Lander, Eric; Gabriel, Stacey B.; Getz, Gad; et al. 2015. "Comprehensive Genomic Characterization of Head and Neck Squamous Cell Carcinomas." *Nature* 517 (7536): 576–82.

Li, Heng; Handsaker, Bob; Wysoker, Alec; Fennell, Tim; Ruan, Jue; Homer, Nils; Marth, Gabor; Abecasis, Goncalo; Durbin, Richard. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Liang, Winnie S; Aldrich, Jessica; Nasser, Sara; Kurdoglu, Ahmet; Phillips, Lori; Reiman, Rebecca; McDonald, Jacquelyn; et al. 2014. "Simultaneous Characterization of Somatic Events and HPV-18 Integration in a Metastatic Cervical Carcinoma Patient Using DNA and RNA Sequencing." *International Journal of Gynecological Cancer* 24 (2): 329–38.

Liang, Y.; Tayo, B.; Cai, X.; Kelemen, A. 2005. "Differential and Trajectory Methods for Time Course Gene Expression Data." *Bioinformatics* 21 (13): 3009–16.

Mardia, K. V.; Kent, J. T.; Bibby, J. M. 1979. *Multivariate Analysis*. London: Academic Press.

McLaughlin-Drubin, Margaret E.; Munger, Karl. 2008. "Viruses Associated with Human Cancer." *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1782 (3): 127–50.

Moore, Patrick S; Chang, Yuan. 2010. "Why Do Viruses Cause Cancer? Highlights of the First Century of Human Tumour Virology." *Nature Reviews Cancer* 10 (12): 878–89.

Morales-Sánchez, Abigail; Fuentes-Pananá, Ezequiel M. 2014. "Human Viruses and Cancer." *Viruses* 6 (10): 4047–79.

Muñoz, Nubia; Bosch, F. Xavier; de Sanjosé, Silvia; Herrero, Rolando; Castellsagué, Xavier; Shah, Keerti V.; Snijders, Peter J.F.; Meijer, Chris J.L.M. 2003. "Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer." *New England Journal of Medicine* 348 (6): 518–27.

Ogata, H; Goto, S; Sato, K; Fujibuchi, W; Bono, H; Kanehisa, M. 1999. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 27 (1): 29–34.

Patel, P S; Shah, M H; Jha, F P; Raval, G N; Rawal, R M; Patel, M M; Patel, J B; Patel, D D. 2004. "Alterations in Plasma Lipid Profile Patterns in Head and Neck Cancer and Oral Precancerous Conditions." *Indian Journal of Cancer* 41 (1): 25–31.

Raju, Kalyani; Punnayanapalya, Shruthi Suresh; Mariyappa, Narayanaswamy; Eshwarappa, Sumathi Mayagondanahalli; Anjaneya, Chandramouli; Jun, Lee. 2014. "Significance of the Plasma Lipid Profile in Cases of Carcinoma of Cervix : A Tertiary Hospital Based Study" 15: 3779–84.

Reimand, Jüri; Arak, Tambet; Adler, Priit; Kolberg, Liis; Reisberg, Sulev; Peterson, Hedi; Vilo, Jaak. 2016. "g:Profiler—a Web Server for Functional Interpretation of Gene Lists (2016 Update)." *Nucleic Acids Research* 44 (April): W83–W89.

Schmieder, Robert; Edwards, Robert. 2011. "Quality Control and Preprocessing of Metagenomic Datasets." *Bioinformatics* 27 (6): 863–64.

Siegel, Abby B.; Zhu, Andrew X. 2009. "Metabolic Syndrome and Hepatocellular Carcinoma." *Cancer* 115 (24): 5651–61.

Singh, Simranjit; Ramesh, Venkatapathy; Premalatha, Balakrishnan; Prashad, KarthikshreeVishnu; Ramadoss, Koliyan. 2013. "Alterations in Serum Lipid Profile Patterns in Oral Cancer." *Journal of Natural Science, Biology and Medicine* 4 (2): 374–78.

Subramanian, Aravind; Kuehn, Heidi; Gould, Joshua; Tamayo, Pablo; Mesirov, Jill P. 2007. "GSEA-P: A Desktop Application for Gene Set Enrichment Analysis." *Bioinformatics* 23 (23): 3251–53.

Subramanian, Aravind; Tamayo, Pablo; Mootha, Vamsi K; Mukherjee, Sayan; Ebert, Benjamin L; Gillette, Michael a; Paulovich, Amanda; et al. 2005. "Gene Set Enrichment Analysis: A

Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.

Tang, Ka-Wei; Alaei-Mahabadi, Babak; Samuelsson, Tore; Lindh, Magnus; Larsson, Erik. 2013. "The Landscape of Viral Expression and Host Gene Fusion and Adaptation in Human Cancer." *Nature Communications* 4 (October). Nature Publishing Group: 2513.

Tomczak, Katarzyna; Czerwińska, Patrycja; Wiznerowicz, Maciej. 2015. "Review The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge." *Współczesna Onkologia* 1A: 68–77.

Yan, Li; Zhan, Cheng; Wu, Jihong; Wang, Shengzi. 2016. "Expression Profile Analysis of Head and Neck Squamous Cell Carcinomas Using Data from The Cancer Genome Atlas." *Molecular Medicine Reports* 13 (5): 4259–65.

Yao, Fangzhou; Coquery, Jeff; Lê Cao, Kim-Anh. 2012. "Independent Principal Component Analysis for Biologically Meaningful Dimension Reduction of Large Biological Data Sets." *BMC Bioinformatics* 13 (February). BioMed Central: 24.

Zhang, Junjun; Baran, Joachim; Cros, A; Guberman, Jonathan M; Haider, Syed; Hsu, Jack; Liang, Yong; et al. 2011. "International Cancer Genome Consortium Data Portal—a One-Stop Shop for Cancer Genomics Data." *Database: The Journal of Biological Databases and Curation* 2011 (September). Oxford University Press: bar026.

Zhang, Wensheng; Edwards, Andrea; Fang, Zhide; Flemington, Erik K; Zhang, Kun. 2016. "Integrative Genomics and Transcriptomics Analysis Reveals Potential Mechanisms for Favorable Prognosis of Patients with HPV-Positive Head and Neck Carcinomas." *Scientific Reports* 6 (April). Nature Publishing Group: 24927.

ANNEXES

Table 1. *CESC resume table of viral infection.*

CESC Samples	Nº Reads	Human Reads	Non-Human Reads	Viral Infection	Virus Found	Total Human Viral Reads	Total ppm of Human Viral Reads	Human Virus with ppm>10	Sum of all Reads with ppm>10	Sum of all ppm bigger than 10	Bigger ppm found	Virus with bigger ppm
TCGA-4J-AA1J-01A-21R-A38B-07	118091872	3351580	121443452	Yes	7	29215	8716.784324	5	29209	8714.994123	6150.830355	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-Q1-A73P-01A-11R-A32P-07	129566280	4144362	133710642	Yes	17	31573	7618.301685	9	31534	7608.89131	3867.66407	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-VS-A9UP-01A-11R-A42T-07	115495726	2593965	118089691	Yes	19	13543	5220.964815	8	13475	5194.750123	3247.923546	NC_001357 » Human papillomavirus - 18
TCGA-VS-A9V5-01A-11R-A42T-07	120683236	2656068	123339304	Yes	10	10807	4068.796432	5	10772	4055.619059	2896.010193	NC_001526 » Human papillomavirus type 16
TCGA-VS-A957-01A-11R-A42T-07	53034359	1359417	54393776	Yes	12	5013	3687.610201	6	4998	3676.576061	2721.754988	M12732 » Human papillomavirus type 33 complete genome
TCGA-ZJ-AAXD-01A-21R-A42T-07	61842201	1609661	63451862	Yes	9	4303	2673.233682	5	4280	2658.944959	2359.503026	NC_001526 » Human papillomavirus type 16
TCGA-C5-A3HD-01B-11R-A213-07	166691923	4492768	171184691	Yes	12	13158	2928.706757	5	13100	2915.797121	2201.983276	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9UL-01A-11R-A42T-07	103948945	2167696	106116641	Yes	16	7747	3573.840612	7	7685	3545.238816	2189.421395	NC_001357 » Human papillomavirus - 18
TCGA-VS-A9V3-01A-11R-A42T-07	83914920	2046142	85961062	Yes	9	6169	3014.942266	5	6132	2996.859455	2150.388389	D90400 » Human papillomavirus type 58 complete genome
TCGA-VS-A9UD-01A-11R-A42T-07	65306360	2004117	67310477	Yes	12	4739	2364.632405	5	4682	2336.190951	2111.154189	NC_001526 » Human papillomavirus type 16
TCGA-Q1-A6DT-01A-11R-A32P-07	122568863	3693586	126262449	Yes	16	11063	2995.192207	5	10995	2976.781912	2102.022262	NC_001526 » Human papillomavirus type 16
TCGA-EA-A439-01A-11R-A24H-07	237540117	7233358	244773475	Yes	18	25534	3530.034044	8	25363	3506.393573	1883.495881	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-IR-A3LB-01A-11R-A24H-07	227676657	6606611	234283268	Yes	12	20520	3105.979754	5	20465	3097.654758	1879.935113	NC_001357 » Human papillomavirus - 18
TCGA-Q1-A73O-01A-11R-A32P-07	160629916	4531268	165161184	Yes	16	15447	3408.979562	7	15356	3388.896882	1865.923622	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-EA-A1QS-01A-61R-A22U-07	176704793	6633869	183338662	Yes	10	21716	3273.504497	4	21593	3254.96328	1862.110934	FR751039 » Human papillomavirus type 68b complete genome
TCGA-EK-A2H0-01A-11R-A180-07	178994152	7175668	186169820	Yes	13	21796	3037.487243	5	21693	3023.133177	1827.704403	M62849 » Human papillomavirus ORFs (HPV-39)
TCGA-VS-A9UI-01A-11R-A42T-07	82890073	2573902	85463975	Yes	9	5147	1999.687634	4	5125	1991.1403	1704.416097	X74481 » Human papillomavirus type 52 genomic DNA
TCGA-VS-AA62-01A-11R-A42T-07	116243429	2910235	119153664	Yes	13	8837	3036.524542	4	8746	3005.25559	1621.862152	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-C5-A3HF-01A-11R-A213-07	169111651	6391918	175503569	Yes	18	20429	3196.067284	10	20373	3187.30622	1604.056873	NC_001357 » Human papillomavirus - 18
TCGA-VS-A9UZ-01A-11R-A42T-07	90519136	2419872	92939008	Yes	11	4455	1841.006467	3	4390	1814.145542	1601.737613	NC_001526 » Human papillomavirus type 16
TCGA-EA-A97N-01A-11R-A38B-07	111480979	2857908	114338887	Yes	8	5355	1873.748209	4	5307	1856.952708	1566.530483	NC_001526 » Human papillomavirus type 16
TCGA-WL-A834-01A-11R-A352-07	112613702	4992324	117606026	Yes	13	8415	1685.587718	4	8330	1668.561576	1540.765383	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9UV-01A-11R-A42T-07	95961856	1895259	97857115	Yes	11	3107	1639.353777	4	3072	1620.886645	1505.335155	X74481 » Human papillomavirus type 52 genomic DNA
TCGA-LP-A4AX-01A-12R-A24H-07	105651053	3353918	109004971	Yes	9	5394	1608.268301	4	5327	1588.291663	1407.011143	NC_001526 » Human papillomavirus type 16
TCGA-VS-A953-01A-11R-A38B-07	177720560	6174986	183895546	Yes	16	10920	1768.425062	5	10773	1744.619339	1362.108351	NC_001526 » Human papillomavirus type 16
TCGA-VS-A8EB-01A-11R-A36F-07	142428353	4006316	146434669	Yes	12	7577	1891.263696	4	7487	1868.799166	1362.848063	NC_001526 » Human papillomavirus type 16

TCGA-VS-A9UB-01A-22R-A42T-07	84908691	1999318	86908009	Yes	10	3936	1968.671319	3	3921	1961.168759	1354.962042	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-EK-A2RM-01A-21R-A18M-07	144276484	4099513	148375997	Yes	15	10585	2582.014009	4	10471	2554.205829	1348.452853	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-FU-A3TX-01A-11R-A22U-07	197613889	4560827	202174716	Yes	15	14619	3205.339733	5	14441	3166.311724	1335.284149	NC_001357 » Human papillomavirus - 18
TCGA-EA-A5FO-01A-21R-A28H-07	113840288	3518377	117358665	Yes	10	6720	1909.971558	6	6692	1902.013343	1332.432539	NC_001526 » Human papillomavirus type 16
TCGA-MY-A913-01A-11R-A37O-07	150316371	5655598	155971969	Yes	14	13231	2339.45199	5	13213	2336.269302	1303.840902	NC_001357 » Human papillomavirus - 18
TCGA-ZJ-AAX8-01A-11R-A42T-07	138013957	3406767	141420724	Yes	9	8403	2466.561405	4	8382	2460.397204	1291.547088	M62849 » Human papillomavirus ORFs (HPV-39)
TCGA-JW-AAVH-01A-11R-A38B-07	118290407	2883947	121174354	Yes	7	4170	1445.935033	3	4134	1433.452141	1275.682251	NC_001526 » Human papillomavirus type 16
TCGA-MA-AA3Y-01A-11R-A38B-07	147969259	6909483	154878742	Yes	12	21870	3165.215111	6	21775	3151.465891	1250.74481	FR751039 » Human papillomavirus type 68b complete genome
TCGA-JW-A5VL-01A-11R-A28H-07	149414711	5222922	154637633	Yes	10	12200	2335.857208	7	12191	2334.134034	1247.768969	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1M9-01A-11R-A13Y-07	216879099	8679070	225558169	Yes	18	15948	1837.524067	6	15862	1827.61517	1245.064275	NC_001526 » Human papillomavirus type 16
TCGA-MA-AA41-01A-11R-A38B-07	143550983	5066468	148617451	Yes	11	8616	1700.592994	5	8523	1682.237013	1236.166892	NC_001526 » Human papillomavirus type 16
TCGA-FU-A40J-01A-11R-A24H-07	195416145	6482615	201898760	Yes	15	10947	1688.670389	7	10880	1678.335054	1234.069893	NC_001526 » Human papillomavirus type 16
TCGA-VS-A8EC-01A-11R-A36F-07	145345989	4502129	149848118	Yes	11	6784	1506.842651	5	6724	1493.515623	1228.974114	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9UC-01A-11R-A42T-07	127093012	3600515	130693527	Yes	11	6250	1735.862786	5	6200	1721.975884	1227.324424	M12732 » Human papillomavirus type 33 complete genome
TCGA-IR-A3LF-01A-21R-A22U-07	164084856	5502722	169587578	Yes	13	8732	1586.851016	5	8643	1570.677204	1215.035032	NC_001526 » Human papillomavirus type 16
TCGA-C5-A2M2-01A-21R-A18M-07	140221573	4746740	144968313	Yes	11	6976	1469.640217	5	6878	1448.994468	1198.506765	NC_001526 » Human papillomavirus type 16
TCGA-VS-A94X-01A-11R-A38B-07	128812958	3336341	132149299	Yes	16	4948	1483.061836	4	4792	1436.304024	1195.621191	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-C5-A7CH-01A-11R-A33Z-07	122764021	4977330	127741351	Yes	11	7233	1453.188757	4	7123	1431.088555	1147.402322	NC_001526 » Human papillomavirus type 16
TCGA-C5-A7X8-01A-11R-A36F-07	119803218	7359797	127163015	Yes	16	17402	2364.467389	6	17279	2347.754972	1144.73266	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-VS-A9UQ-01A-21R-A42T-07	114021267	2987841	117009108	Yes	9	3818	1277.845777	2	3758	1257.764386	1134.263838	NC_001526 » Human papillomavirus type 16
TCGA-VS-A959-01A-11R-A42T-07	88133059	2873312	91006371	Yes	12	3594	1250.82135	2	3555	1237.248165	1134.231159	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-Q1-A5R1-01A-11R-A28H-07	113568275	5393780	118962055	Yes	12	8103	1502.285966	5	8014	1485.785479	1101.639296	NC_001526 » Human papillomavirus type 16
TCGA-Q1-A73R-01A-11R-A33Z-07	188229981	8578467	196808448	Yes	18	11706	1364.579478	5	11489	1339.283582	1058.114463	NC_001526 » Human papillomavirus type 16
TCGA-C5-A2LZ-01A-11R-A213-07	173570183	5738326	179308509	Yes	10	7866	1370.783047	3	7699	1341.680483	1055.882848	NC_001526 » Human papillomavirus type 16
TCGA-EX-A69L-01A-11R-A32P-07	131369701	5000928	136370629	Yes	11	5994	1198.577545	5	5965	1192.778621	1025.009758	NC_001526 » Human papillomavirus type 16
TCGA-C5-A3HL-01A-11R-A213-07	178906142	5548288	184454430	Yes	15	7526	1356.45446	4	7393	1332.483101	1025.361337	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9U7-01A-11R-A42T-07	207451337	4164346	211615683	Yes	12	5514	1324.097469	4	5437	1305.607171	1022.969753	NC_001526 » Human papillomavirus type 16
TCGA-EX-A449-01A-11R-A32Y-07	133858274	6361710	140219984	Yes	15	8786	1381.07521	6	8745	1374.630405	1020.794723	NC_001526 » Human papillomavirus type 16
TCGA-C5-A7CJ-01A-11R-A32P-07	129823759	4172220	133995979	Yes	12	4944	1184.980658	3	4846	1161.491963	1016.724909	NC_001526 » Human papillomavirus type 16
TCGA-C5-A907-01A-11R-A37O-07	116969582	4949388	121918970	Yes	12	8155	1647.67846	2	7995	1615.351231	1014.87295	NC_001357 » Human papillomavirus - 18
TCGA-ZJ-AAXN-01A-11R-A42T-07	247314173	9387967	256702140	Yes	33	18650	1986.58559	6	18449	1965.175207	1007.459869	NC_001357 » Human papillomavirus - 18
TCGA-ZJ-A8QR-01A-11R-A37O-07	113212977	5039842	118252819	Yes	11	5975	1185.553039	3	5878	1166.306404	993.880364	NC_001526 » Human papillomavirus type 16
TCGA-DS-A1OC-01A-11R-A14Y-07	126179430	5083017	131262447	Yes	13	5461	1074.361941	3	5384	1059.213456	969.896422	NC_001526 » Human papillomavirus type 16

TCGA-C5-A1MJ-01A-11R-A14Y-07	201716763	9304831	211021594	Yes	20	19955	2144.584894	6	19772	2124.917692	964.767657	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-JX-A3Q0-01A-11R-A32Y-07	146493850	4919521	151413371	Yes	16	5949	1209.264072	4	5754	1169.626067	957.410284	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-ZJ-AAXI-01A-11R-A42T-07	112611559	4506206	117117765	Yes	11	5021	1114.241115	3	4915	1090.718001	950.245062	NC_001526 » Human papillomavirus type 16
TCGA-EK-A3GN-01A-11R-A213-07	198451487	8044708	206496195	Yes	17	11031	1371.211979	5	10869	1351.07452	945.839178	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1ML-01A-11R-A14Y-07	107424587	5327311	112751898	Yes	7	5708	1071.45988	3	5617	1054.378091	945.505153	NC_001526 » Human papillomavirus type 16
TCGA-DS-A0VK-01A-21R-A10U-07	239743463	5491184	245234647	Yes	13	6434	1171.696304	3	6313	1149.660984	943.876585	NC_001526 » Human papillomavirus type 16
TCGA-LP-A5U2-01A-11R-A28H-07	141366205	4080020	145446225	Yes	11	4319	1058.573241	4	4236	1038.230204	925.975853	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9U5-01A-11R-A42T-07	76738412	2350578	79088990	Yes	7	2327	989.969275	2	2299	978.057312	913.817793	NC_001526 » Human papillomavirus type 16
TCGA-EA-A3HS-01A-11R-A213-07	202725170	6151663	208876833	Yes	8	6835	1111.081669	3	6700	1089.136385	900.894604	NC_001526 » Human papillomavirus type 16
TCGA-ZJ-A8QO-01A-11R-A37O-07	148839160	6155499	154994659	Yes	18	9454	1535.862483	5	9352	1519.291936	897.246511	M12732 » Human papillomavirus type 33 complete genome
TCGA-VS-A94W-01A-12R-A37O-07	143424338	6468055	149892393	Yes	13	7478	1156.14354	3	7273	1124.449313	896.405488	NC_001526 » Human papillomavirus type 16
TCGA-ZJ-A8QQ-01A-11R-A37O-07	139944411	9235197	149179608	Yes	14	14183	1535.755004	3	13929	1508.251529	892.780089	M62849 » Human papillomavirus ORFs (HPV-39)
TCGA-C5-A7UH-01A-11R-A352-07	140147683	5956000	146103683	Yes	12	6819	1144.895904	3	6666	1119.207522	892.377435	NC_001526 » Human papillomavirus type 16
TCGA-JX-A5QV-01A-22R-A28H-07	134797908	5574000	140371908	Yes	10	5718	1025.834229	4	5637	1011.302476	891.998565	NC_001526 » Human papillomavirus type 16
TCGA-C5-A8XJ-01A-11R-A37O-07	132756585	4849741	137606326	Yes	17	6198	1278.006394	4	6108	1259.4487	889.325842	M12732 » Human papillomavirus type 33 complete genome
TCGA-MA-AA3W-01A-11R-A38B-07	112777715	3670756	116448471	Yes	7	3664	998.159507	4	3651	994.618002	874.206839	NC_001526 » Human papillomavirus type 16
TCGA-HM-A4S6-01A-11R-A26T-07	129184613	4479509	133664122	Yes	11	4431	989.170914	3	4365	974.437154	860.58539	NC_001526 » Human papillomavirus type 16
TCGA-EA-A411-01A-11R-A24H-07	231960683	7132378	239093061	Yes	15	7912	1109.307444	4	7802	1093.884816	838.990867	NC_001526 » Human papillomavirus type 16
TCGA-C5-A7CK-01A-11R-A32P-07	174197658	4344085	178541743	Yes	12	4577	1053.616583	3	4527	1042.106681	837.460593	NC_001526 » Human papillomavirus type 16
TCGA-VS-A8EL-01A-11R-A37O-07	146339443	6092782	152432225	Yes	13	6244	1024.819204	3	6094	1000.199908	828.028969	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-DS-A7WF-01A-11R-A352-07	143071789	6932643	150004432	Yes	11	6582	949.421454	3	6542	943.651649	822.918474	NC_001526 » Human papillomavirus type 16
TCGA-C5-A7CG-01A-11R-A32P-07	144031929	4319414	148351343	Yes	10	4444	1028.843265	4	4433	1026.296623	821.870745	X74481 » Human papillomavirus type 52 genomic DNA
TCGA-JX-A3Q8-01A-11R-A21T-07	157739769	5717823	163457592	Yes	12	6309	1103.391972	4	6248	1092.723577	815.170389	NC_001526 » Human papillomavirus type 16
TCGA-Q1-A6DV-01A-11R-A32P-07	134257054	5919744	140176798	Yes	11	5869	991.428008	3	5791	978.251762	812.872989	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1MK-01A-11R-A14Y-07	151353537	7239505	158593042	Yes	13	9916	1369.706908	4	9793	1352.716795	801.021617	D90400 » Human papillomavirus type 58 complete genome
TCGA-EX-A1H5-01A-31R-A13Y-07	176326973	5544618	181871591	Yes	10	5530	997.363569	3	5424	978.245931	794.464109	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9UO-01A-11R-A42T-07	204733331	5135786	209869117	Yes	15	5738	1117.258388	4	5642	1098.566022	784.690016	NC_001357 » Human papillomavirus - 18
TCGA-C5-A2LS-01A-22R-A22U-07	136158371	5431129	141589500	Yes	9	4942	909.939721	3	4896	901.470026	780.86895	NC_001526 » Human papillomavirus type 16
TCGA-Q1-A73Q-01A-21R-A32P-07	127836430	11688173	139524603	Yes	26	12560	1074.590531	6	12303	1052.602489	759.485678	M12732 » Human papillomavirus type 33 complete genome
TCGA-ZJ-AB0H-01A-11R-A42T-07	95544488	3309924	98854412	Yes	11	4462	1348.067204	4	4374	1321.480493	750.772525	NC_001357 » Human papillomavirus - 18
TCGA-EA-A3QD-01A-32R-A22U-07	158959506	8457937	167417443	Yes	9	8232	973.286985	5	8135	961.818467	749.473542	NC_001526 » Human papillomavirus type 16
TCGA-C5-A8XH-01A-11R-A37O-07	169848532	5834920	175683452	Yes	12	5129	879.01805	3	5047	864.964729	742.426631	NC_001526 » Human papillomavirus type 16
TCGA-Q1-A6DW-01A-11R-A32P-07	139944972	3883024	143827996	Yes	8	5177	1333.239247	2	5140	1323.710592	733.706513	J04353 » Human papillomavirus type 31 (HPV-31) complete genome

TCGA-EK-A2R8-01A-21R-A18M-07	136482168	7475828	143957996	Yes	19	10726	1434.757458	3	10494	1403.724109	724.468246	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-ZJ-AAXA-01A-11R-A42T-07	272523092	6193407	278716499	Yes	13	5673	915.974036	4	5598	903.864384	721.573764	NC_001526 » Human papillomavirus type 16
TCGA-MA-AA3X-01A-22R-A42S-07	124599784	8329922	132929706	Yes	11	8522	1023.058799	4	8385	1006.612066	721.375302	NC_001526 » Human papillomavirus type 16
TCGA-DG-A2KL-01A-11R-A180-07	178951244	6826226	185777470	Yes	13	6100	893.612372	3	5934	869.294395	721.628613	NC_001526 » Human papillomavirus type 16
TCGA-JX-A3PZ-01A-11R-A32Y-07	145232552	5228089	150460641	Yes	17	6083	1163.522653	3	5986	1144.969032	713.262532	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-VS-A9V2-01A-11R-A42T-07	101349477	2563397	103912874	Yes	6	2035	793.86845	3	2026	790.357483	711.945906	NC_001526 » Human papillomavirus type 16
TCGA-VS-A8QF-01A-21R-A37O-07	138574255	6238983	144813238	Yes	12	6368	1020.679174	3	6238	999.842443	707.166537	NC_001526 » Human papillomavirus type 16
TCGA-FU-A2QG-01A-11R-A18M-07	201951071	8084747	210035818	Yes	15	7587	938.433821	3	7430	919.014535	706.639305	NC_001526 » Human papillomavirus type 16
TCGA-EA-A43B-01A-81R-A32Y-07	101403793	3976917	105380710	Yes	8	3196	803.63759	3	3146	791.065038	703.81152	NC_001526 » Human papillomavirus type 16
TCGA-EA-A3HQ-01A-11R-A213-07	229193607	9788167	238981774	Yes	12	10248	1046.978459	5	10141	1036.046893	703.196012	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9V1-01A-11R-A42T-07	78251056	2151137	80402193	Yes	5	2226	1034.801595	3	2222	1032.942113	697.305657	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-LP-A5U3-01A-11R-A28H-07	128040465	5523660	133564125	Yes	7	4320	782.090136	2	4279	774.667521	693.923956	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9UY-01A-11R-A42T-07	116596054	2535098	119131152	Yes	9	2221	876.100253	3	2181	860.321771	692.28093	NC_001526 » Human papillomavirus type 16
TCGA-EA-A1QT-01A-11R-A14Y-07	244524044	9711902	254235946	Yes	15	9064	933.287834	3	8823	908.472923	692.34636	NC_001526 » Human papillomavirus type 16
TCGA-DS-A5RQ-01A-11R-A28H-07	122779875	4125668	126905543	Yes	11	3267	791.871765	3	3217	779.752515	688.373374	NC_001526 » Human papillomavirus type 16
TCGA-C5-A7CO-01A-11R-A352-07	145507646	6231500	151739146	Yes	13	4487	720.051352	2	4426	710.262377	681.858301	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-ZJ-AAXB-01A-11R-A42T-07	61516995	1612059	63129054	Yes	3	1282	795.256253	2	1281	794.635928	678.635211	NC_001357 » Human papillomavirus - 18
TCGA-MY-A5BF-01A-11R-A26T-07	124490541	5413223	129903764	Yes	9	4365	806.3588	3	4304	795.090097	674.090094	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1MP-01A-11R-A14Y-07	131469235	5966041	137435276	Yes	11	4597	770.527724	2	4444	744.882578	674.65175	NC_001526 » Human papillomavirus type 16
TCGA-JW-A5VG-01A-11R-A28H-07	141858852	5521838	147380690	Yes	13	3948	714.979323	3	3860	699.042602	673.869824	X77858 » Human papilloma virus type 59 complete viral genome
TCGA-EA-A3HU-01A-11R-A213-07	197792377	7491877	205284254	Yes	13	6426	857.728979	3	6302	841.177718	668.323839	NC_001526 » Human papillomavirus type 16
TCGA-JW-A69B-01A-11R-A32P-07	178359672	6159869	184519541	Yes	10	4846	786.705041	3	4794	778.263304	667.871346	NC_001526 » Human papillomavirus type 16
TCGA-EA-A5O9-01A-11R-A28H-07	140000706	4285092	144285798	Yes	7	3346	780.846711	4	3343	780.14661	666.496775	NC_001526 » Human papillomavirus type 16
TCGA-DS-A0VM-01A-11R-A10U-07	200911285	9632137	210543422	Yes	15	7827	812.592261	3	7635	792.658991	659.147601	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9UR-01A-11R-A42T-07	103100263	2786810	105887073	Yes	6	2106	755.702756	3	2091	750.320258	655.947122	NC_001526 » Human papillomavirus type 16
TCGA-C5-A902-01A-11R-A37O-07	140473634	8012075	148485709	Yes	13	6539	816.143133	3	6380	796.298088	650.268501	NC_001526 » Human papillomavirus type 16
TCGA-EK-A2PI-01A-11R-A18M-07	189539351	5925127	195464478	Yes	11	4692	791.881761	2	4533	765.046893	642.855419	NC_001526 » Human papillomavirus type 16
TCGA-EA-A3HT-01A-61R-A21T-07	199406770	8506245	207913015	Yes	14	6831	803.05705	3	6654	782.248807	631.065764	NC_001526 » Human papillomavirus type 16
TCGA-FU-A3NI-01A-11R-A21T-07	216359088	7928957	224288045	Yes	11	6045	762.395358	3	5906	744.864678	629.843244	NC_001526 » Human papillomavirus type 16
TCGA-FU-A770-01A-11R-A33Z-07	110252004	8154653	118406657	Yes	13	8573	1051.301631	4	8493	1041.491282	621.117784	NC_001526 » Human papillomavirus type 16
TCGA-EK-A3GK-01A-11R-A213-07	189910737	8096855	198007592	Yes	10	5990	739.793413	3	5876	725.713873	613.571566	NC_001526 » Human papillomavirus type 16
TCGA-DG-A2KH-01A-21R-A22U-07	203504157	7329666	210833823	Yes	13	10947	1493.519624	3	10852	1480.558596	613.806959	NC_001357 » Human papillomavirus - 18
TCGA-IR-A3LI-01A-11R-A32Y-07	126451347	4588136	131039483	Yes	10	3064	667.809322	3	3024	659.091187	611.577338	NC_001526 » Human papillomavirus type 16

TCGA-VS-A952-01A-11R-A38B-07	123681770	7268306	130950076	Yes	12	5194	714.609429	2	5063	696.585972	606.331104	NC_001526 » Human papillomavirus type 16
TCGA-VS-A8QA-01A-11R-A37O-07	127394786	5373674	132768460	Yes	14	3858	717.944555	3	3806	708.267751	600.706332	NC_001526 » Human papillomavirus type 16
TCGA-DG-A2KK-01A-11R-A180-07	173549755	7375098	180924853	Yes	6	6109	828.327975	4	6045	819.650124	595.110736	NC_001526 » Human papillomavirus type 16
TCGA-ZJ-AB0I-01A-11R-A42T-07	103149568	3273050	106422618	Yes	8	2285	698.125601	2	2256	689.265364	592.108278	NC_001526 » Human papillomavirus type 16
TCGA-ZJ-AAXJ-01A-11R-A42T-07	103915425	3255116	107170541	Yes	8	2060	632.849951	2	2005	615.953471	591.683983	NC_001526 » Human papillomavirus type 16
TCGA-LP-A7HU-01A-11R-A33Z-07	130799586	9723303	140522889	Yes	13	6638	682.689823	4	6527	671.273949	591.671369	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9UM-01A-11R-A42T-07	101668127	2309241	103977368	Yes	6	1437	622.282387	2	1428	618.385002	585.473755	NC_001526 » Human papillomavirus type 16
TCGA-BI-A20A-01A-11R-A14Y-07	164775983	9189583	173965566	Yes	14	6410	697.528931	3	6244	679.464999	577.828178	NC_001526 » Human papillomavirus type 16
TCGA-MA-AA42-01A-12R-A38B-07	111198721	4840178	116038899	Yes	7	3173	655.554404	2	3135	647.703452	572.499606	NC_001526 » Human papillomavirus type 16
TCGA-C5-A8XK-01A-11R-A37O-07	128487978	7941438	136429416	Yes	11	5587	703.52498	3	5464	688.0366	567.781301	NC_001526 » Human papillomavirus type 16
TCGA-C5-A2M1-01A-11R-A18M-07	257073020	9478631	266551651	Yes	15	6908	728.797227	3	6731	710.123646	561.578988	NC_001526 » Human papillomavirus type 16
TCGA-DS-A3LQ-01A-21R-A21T-07	199898973	6311125	206210098	Yes	8	5966	945.31482	2	5915	937.233853	558.379053	NC_001583 » Human papillomavirus type 26
TCGA-EA-A6QX-01A-12R-A33Z-07	128494146	7283625	135777771	Yes	16	6947	953.783317	4	6821	936.484238	551.785684	D90400 » Human papillomavirus type 58 complete genome
TCGA-C5-A2LX-01A-11R-A18M-07	188562235	15620784	204183019	Yes	16	10024	641.709147	3	9663	618.598913	525.517797	NC_001526 » Human papillomavirus type 16
TCGA-C5-A905-01A-11R-A37O-07	145144616	7592161	152736777	Yes	11	4809	633.416495	2	4685	617.083858	516.980607	NC_001526 » Human papillomavirus type 16
TCGA-C5-A901-01A-11R-A37O-07	149847526	5850883	155698409	Yes	9	3686	629.990378	3	3641	622.299232	510.521232	NC_001526 » Human papillomavirus type 16
TCGA-IR-A3LL-01A-11R-A213-07	186899012	7669114	194568126	Yes	10	4507	587.681966	3	4446	579.727985	502.926414	NC_001526 » Human papillomavirus type 16
TCGA-C5-A7X3-01A-11R-A352-07	117566394	8184756	125751150	Yes	10	6756	825.436946	5	6745	824.092985	501.419957	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-C5-A1MN-01A-11R-A14Y-07	140085948	9583208	149669156	Yes	11	5655	590.094673	3	5575	581.746739	499.10218	NC_001526 » Human papillomavirus type 16
TCGA-XS-A8TJ-01A-11R-A36F-07	147186898	5531608	152718506	Yes	10	3033	548.303495	2	2948	532.937258	493.708159	NC_001526 » Human papillomavirus type 16
TCGA-VS-A94Z-01A-11R-A38B-07	133732493	3192282	136924775	Yes	8	2444	765.596524	3	2423	759.018157	487.738865	NC_001526 » Human papillomavirus type 16
TCGA-EK-A2R9-01A-11R-A18M-07	166223238	9197301	175420539	Yes	12	8386	911.789232	3	8222	893.957913	484.489961	M12732 » Human papillomavirus type 33 complete genome
TCGA-C5-A1MI-01A-11R-A14Y-07	220557148	8700193	229257341	Yes	16	7604	874.0036	3	7290	837.912447	480.793932	NC_001357 » Human papillomavirus - 18
TCGA-MA-AA3Z-01A-11R-A38B-07	104187538	3213917	107401455	Yes	9	1812	563.798008	2	1764	548.862961	475.743462	NC_001526 » Human papillomavirus type 16
TCGA-FU-A23K-01A-11R-A16R-07	182868672	8597004	191465676	Yes	17	5436	632.313306	4	5391	627.078923	474.816576	NC_001357 » Human papillomavirus - 18
TCGA-FU-A3TQ-01A-11R-A22U-07	207605213	10152439	217757652	Yes	13	5820	573.261259	3	5701	561.539942	471.413815	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1BQ-01C-11R-A213-07	200148840	10987374	211136214	Yes	17	6097	554.909666	2	5973	543.623982	470.085027	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-MY-A5BD-01A-11R-A26T-07	120065637	4537862	124603499	Yes	9	2354	518.746492	3	2336	514.779867	466.959991	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9UU-01A-11R-A42T-07	75284823	2171334	77456157	Yes	7	1146	527.786143	2	1127	519.035763	461.928013	NC_001526 » Human papillomavirus type 16
TCGA-DS-A1OD-01A-11R-A14Y-07	246343358	12105033	258448391	Yes	15	5961	492.439798	3	5863	484.343992	452.125988	NC_001526 » Human papillomavirus type 16
TCGA-JW-A5VK-01A-11R-A28H-07	141447515	4625224	146072739	Yes	8	3373	729.261977	3	3364	727.316126	446.032452	AB027020 » Human papillomavirus type 69 DNA complete genome
TCGA-EK-A2H1-01A-11R-A180-07	147079197	4917541	151996738	Yes	8	2432	494.556121	2	2386	485.201852	445.344533	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9U6-01A-11R-A42T-07	157350502	3821938	161172440	Yes	11	2469	646.00734	4	2389	625.075551	439.305923	M12732 » Human papillomavirus type 33 complete genome

TCGA-VS-A8QM-01A-11R-A37O-07	139117902	9461304	148579206	Yes	12	4739	500.882333	3	4687	495.386259	430.278955	NC_001526 » Human papillomavirus type 16
TCGA-C5-A8XI-01A-11R-A37O-07	149193925	10720112	159914037	Yes	8	5387	502.513408	4	5365	500.46119	429.939538	X74481 » Human papillomavirus type 52 genomic DNA
TCGA-IR-A3L7-01A-21R-A213-07	176435501	5919332	182354833	Yes	13	4215	712.073592	2	4108	693.997228	424.034334	NC_001357 » Human papillomavirus - 18
TCGA-R2-A69V-01A-11R-A32P-07	100015722	3750580	103766302	Yes	10	3348	892.661934	5	3319	884.929797	421.801428	M62849 » Human papillomavirus ORFs (HPV-39)
TCGA-DS-A1OA-01A-11R-A14Y-07	158677679	11770495	170448174	Yes	12	8118	689.690617	4	8024	681.704549	419.268688	D90400 » Human papillomavirus type 58 complete genome
TCGA-EA-A5ZF-01A-11R-A28H-07	128333241	3914623	132247864	Yes	5	2364	603.889569	2	2348	599.802331	411.022977	NC_001357 » Human papillomavirus - 18
TCGA-FU-A3HY-01A-11R-A21T-07	183231438	6334289	189565727	Yes	7	4974	785.249931	3	4962	783.35548	408.096315	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-C5-A2LT-01A-11R-A18M-07	169736075	5970895	175706970	Yes	7	4535	759.517627	3	4489	751.813589	407.979038	U21941 » Human papillomavirus type 70 complete genome
TCGA-Q1-A5R2-01A-11R-A28H-07	141525574	4993423	146518997	Yes	7	3227	646.250078	3	3165	633.833745	404.732385	NC_001526 » Human papillomavirus type 16
TCGA-C5-A3HE-01A-21R-A22U-07	173786860	8984740	182771600	Yes	21	4584	510.198403	2	4332	482.150847	401.569773	NC_001357 » Human papillomavirus - 18
TCGA-EK-A2IP-01A-11R-A180-07	174581621	9326879	183908500	Yes	11	5150	552.167557	4	5053	541.767508	400.777152	NC_001526 » Human papillomavirus type 16
TCGA-EA-A3QE-01A-21R-A21T-07	225645335	12201886	237847221	Yes	10	5839	478.532583	2	5731	469.681491	398.299083	NC_001526 » Human papillomavirus type 16
TCGA-MA-AA43-01A-11R-A42T-07	128954678	2779136	131733814	Yes	7	1287	463.093565	2	1272	457.696205	394.36717	NC_001357 » Human papillomavirus - 18
TCGA-C5-A7CL-01A-11R-A32P-07	136972015	6500033	143472048	Yes	11	3361	517.074296	2	3295	506.920503	390.921092	NC_001526 » Human papillomavirus type 16
TCGA-C5-A7XC-01A-11R-A38B-07	112880915	3502446	116383361	Yes	7	1513	431.983821	2	1499	427.986613	378.021531	NC_001526 » Human papillomavirus type 16
TCGA-LP-A4AU-01A-32R-A32Y-07	151113995	5936720	157050715	Yes	9	3687	621.050007	2	3659	616.333598	375.796736	NC_001357 » Human papillomavirus - 18
TCGA-UC-A7PI-01A-11R-A42S-07	136347452	10296605	146644057	Yes	14	5787	562.02991	2	5744	557.853778	371.09319	NC_001357 » Human papillomavirus - 18
TCGA-ZJ-AAX4-01A-11R-A42T-07	91180487	2390679	93571166	Yes	4	924	386.501073	2	921	385.2462	364.33164	NC_001526 » Human papillomavirus type 16
TCGA-PN-A8MA-01A-11R-A36F-07	123323113	5627038	128950151	Yes	6	2259	401.454547	2	2184	388.126044	360.758182	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1M7-01A-11R-A13Y-07	206481328	7196021	213677349	Yes	17	3461	480.960242	2	3326	462.199874	360.338026	NC_001357 » Human papillomavirus - 18
TCGA-ZJ-AAXF-01A-31R-A42T-07	91244898	2611781	93856679	Yes	4	1035	396.28131	2	1033	395.51555	359.524784	NC_001526 » Human papillomavirus type 16
TCGA-EK-A2RA-01A-11R-A18M-07	188202959	9195352	197398311	Yes	12	3823	415.753526	3	3697	402.05095	357.136954	X74481 » Human papillomavirus type 52 genomic DNA
TCGA-C5-A1M5-01A-11R-A13Y-07	181362670	19774658	201137328	Yes	19	14113	713.691234	3	13662	690.884262	356.82033	D90400 » Human papillomavirus type 58 complete genome
TCGA-VS-A8EI-01A-11R-A37O-07	126431967	11064593	137496560	Yes	10	4094	370.009091	2	4020	363.321091	352.74682	NC_001526 » Human papillomavirus type 16
TCGA-FU-A23L-01A-11R-A16R-07	218138631	9242271	227380902	Yes	11	6072	656.981387	2	6015	650.81407	342.664698	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-UC-A7PF-01A-11R-A352-07	118149410	8562019	126711429	Yes	11	3541	413.570677	2	3450	402.942343	338.705158	NC_001526 » Human papillomavirus type 16
TCGA-VS-A94Y-01A-11R-A38B-07	93585260	3396874	96982134	Yes	5	1365	401.840045	2	1335	393.008395	337.957781	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-VS-A958-01A-11R-A42T-07	95205558	2655211	97860769	Yes	6	1809	681.301787	3	1782	671.133103	334.06008	M62849 » Human papillomavirus ORFs (HPV-39)
TCGA-C5-A7UC-01A-11R-A352-07	130629762	5770563	136400325	Yes	11	2461	426.474852	2	2411	417.810186	325.444848	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-C5-A0TN-01A-21R-A14Y-07	118406550	8400390	126806940	Yes	9	3087	367.482936	2	2971	353.674055	320.818438	NC_001526 » Human papillomavirus type 16
TCGA-MU-A8JM-01A-11R-A36F-07	147839793	5490279	153330072	Yes	5	3368	613.447877	2	3364	612.719317	314.555963	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-C5-A8YR-01A-12R-A37O-07	112474734	7039936	119514670	Yes	9	4077	579.124582	2	4052	575.573414	314.491495	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-C5-A1BL-01A-11R-A13Y-07	130738978	13498688	144237666	Yes	13	4608	341.366509	2	4536	336.032657	313.808275	NC_001526 » Human papillomavirus type 16

TCGA-EX-A3L1-01A-11R-A32Y-07	154269265	6001122	160270387	Yes	9	2721	453.415214	2	2685	447.416333	308.608957	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-DS-A7WH-01A-22R-A352-07	128826930	7559150	136386080	Yes	7	3393	448.859991	3	3378	446.875641	299.372284	NC_001526 » Human papillomavirus type 16
TCGA-C5-A8ZZ-01A-11R-A37O-07	152104547	4640856	156745403	Yes	7	1507	324.724575	2	1472	317.182864	298.6518	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1MQ-01A-11R-A14Y-07	128008947	9928543	137937490	Yes	8	4103	413.252983	3	4054	408.317716	294.403721	NC_001357 » Human papillomavirus - 18
TCGA-Q1-A73S-01A-11R-A33Z-07	100147387	9649617	109797004	Yes	10	4063	421.05298	3	4023	416.907738	288.923384	NC_001357 » Human papillomavirus - 18
TCGA-EA-A5ZD-01A-11R-A28H-07	126246100	4184548	130430648	Yes	11	1502	358.939603	3	1458	348.424728	288.681119	D90400 » Human papillomavirus type 58 complete genome
TCGA-GH-A9DA-01A-21R-A37O-07	142835825	5284945	148120770	Yes	13	2123	401.707113	2	2072	392.05706	286.474126	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-FU-A3YQ-01A-11R-A22U-07	192293156	8858082	201151238	Yes	7	3666	413.859344	2	3617	408.327672	281.212118	NC_001526 » Human papillomavirus type 16
TCGA-EK-A2RB-01A-11R-A18M-07	183938926	10810339	194749265	Yes	12	5389	498.504255	2	5202	481.206001	280.379736	NC_001526 » Human papillomavirus type 16
TCGA-EK-A2GZ-01A-11R-A180-07	213815192	9209927	223025119	Yes	13	2695	292.619038	1	2569	278.93815	278.93815	X74481 » Human papillomavirus type 52 genomic DNA
TCGA-LP-A4AV-01A-11R-A32Y-07	100684338	6377521	107061859	Yes	10	1891	296.510197	2	1838	288.199757	277.850908	X74483 » Human papillomavirus type 56 genomic DNA
TCGA-FU-A3WB-01A-11R-A22U-07	195738966	7817589	203556555	Yes	7	3253	416.112947	2	3200	409.333363	275.14877	NC_001526 » Human papillomavirus type 16
TCGA-IR-A3LK-01A-12R-A213-07	167911447	10437538	178348985	Yes	10	4238	406.034448	2	4073	390.226124	273.819362	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-ZJ-AAXT-01A-11R-A42T-07	137800019	3531925	141331944	Yes	5	1091	308.896707	2	1062	300.685887	267.842607	NC_001357 » Human papillomavirus - 18
TCGA-LP-A4AW-01A-11R-A24H-07	126893348	3522974	130416322	Yes	5	994	282.147981	2	990	281.012577	265.684618	NC_001526 » Human papillomavirus type 16
TCGA-MU-A5YI-01A-11R-A32P-07	131405697	3894952	135300649	Yes	6	1129	289.862367	2	1123	288.32191	264.188108	NC_001526 » Human papillomavirus type 16
TCGA-EK-A2RL-01A-11R-A18M-07	211682677	13387460	225070137	Yes	13	5139	383.866697	3	4987	372.512785	263.679593	NC_001526 » Human papillomavirus type 16
TCGA-EX-A69M-01A-11R-A32P-07	131647198	7196629	138843827	Yes	7	2039	283.327097	2	1975	274.434044	262.345051	X74481 » Human papillomavirus type 52 genomic DNA
TCGA-FU-A5XV-01A-11R-A28H-07	129798235	4411181	134209416	Yes	3	1306	296.065839	2	1305	295.839142	252.313383	NC_001357 » Human papillomavirus - 18
TCGA-EK-A2RK-01A-11R-A18M-07	189028099	7777456	196805555	Yes	13	2305	296.369402	2	2180	280.29731	252.396156	NC_001526 » Human papillomavirus type 16
TCGA-EK-A3GM-01A-11R-A213-07	193731898	8395027	202126925	Yes	11	3537	421.320859	3	3519	419.176734	250.148094	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-EK-A2RJ-01A-11R-A18M-07	178181216	8111669	186292885	Yes	15	2144	264.310586	1	2033	250.626597	250.626597	X77858 » Human papilloma virus type 59 complete viral genome
TCGA-HG-A2PA-01A-11R-A213-07	201003658	6878271	207881929	Yes	6	2055	298.766941	2	2008	291.93383	248.754374	NC_001526 » Human papillomavirus type 16
TCGA-VS-A8EH-01A-11R-A36F-07	152616659	7336929	159953588	Yes	9	2125	289.630717	2	2100	286.223296	245.334254	NC_001526 » Human papillomavirus type 16
TCGA-EK-A2IR-01A-11R-A180-07	232294515	14179130	246473645	Yes	14	6004	423.439236	4	5891	415.469778	240.141673	NC_001526 » Human papillomavirus type 16
TCGA-MU-A51Y-01A-11R-A26T-07	132992611	5101955	138094566	Yes	6	1586	310.86123	2	1578	309.293202	233.439926	NC_001526 » Human papillomavirus type 16
TCGA-VS-A8QC-01A-11R-A37O-07	146963815	6444844	153408659	Yes	7	1760	273.086516	2	1746	270.914237	232.899353	NC_001526 » Human papillomavirus type 16
TCGA-HM-A3JJ-01A-21R-A21T-07	195752623	6840736	202593359	Yes	10	1859	271.754383	2	1845	269.707821	232.577313	NC_001526 » Human papillomavirus type 16
TCGA-C5-A2LV-01A-11R-A18M-07	190662605	8378690	199041295	Yes	13	3046	363.541315	2	2970	354.470687	228.914067	NC_001526 » Human papillomavirus type 16
TCGA-EK-A2R7-01A-11R-A18M-07	188760664	10214792	198975456	Yes	10	2976	291.3422	2	2931	286.936826	226.142637	NC_001357 » Human papillomavirus - 18
TCGA-EK-A2RN-01A-12R-A213-07	172241768	12469940	184711708	Yes	10	3097	248.357249	3	3060	245.390114	218.445317	NC_001526 » Human papillomavirus type 16
TCGA-JW-A5VJ-01A-11R-A28H-07	156471940	5187985	161659925	Yes	8	1325	255.397808	2	1317	253.855784	215.690678	NC_001357 » Human papillomavirus - 18
TCGA-EK-A2PK-01A-11R-A18M-07	148826873	14823054	163649927	Yes	11	4736	319.502309	2	4510	304.25579	213.856065	NC_001357 » Human papillomavirus - 18

TCGA-EK-A2RC-01A-11R-A18M-07	155398617	5833697	161232314	Yes	5	1442	247.18459	2	1402	240.327874	211.015416	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1BE-01B-11R-A13Y-07	192138983	6264610	198403593	Yes	8	1616	257.957001	2	1600	255.40297	207.993794	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1BI-01B-11R-A13Y-07	210804654	35274409	246079063	Yes	19	10791	305.915827	3	10329	292.818514	205.531438	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1BN-01B-11R-A14Y-07	155657974	13983396	169641370	Yes	14	4180	298.925953	2	3988	285.195384	204.814338	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-DS-A1OB-01A-11R-A14Y-07	189816694	9419457	199236151	Yes	9	1979	210.097035	1	1915	203.30259	203.30259	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-EK-A2PG-01A-11R-A18M-07	165959900	26843603	192803503	Yes	17	12300	458.209726	4	12062	449.343555	199.675133	D90400 » Human papillomavirus type 58 complete genome
TCGA-C5-A1MH-01A-11R-A14Y-07	181002911	10135396	191138307	Yes	11	2546	251.198869	2	2507	247.350967	198.709552	NC_001526 » Human papillomavirus type 16
TCGA-DS-A7WI-01A-12R-A352-07	121348063	8872561	130220624	Yes	6	2207	248.744416	2	2178	245.475912	194.419627	NC_001526 » Human papillomavirus type 16
TCGA-DS-A0VN-01A-21R-A10U-07	206453363	28629328	235082691	Yes	14	7543	263.471079	2	6910	241.360887	192.250408	NC_001526 » Human papillomavirus type 16
TCGA-EA-A5ZE-01A-11R-A28H-07	124926194	4340803	129266997	Yes	5	846	194.894815	1	827	190.517745	190.517745	X77858 » Human papilloma virus type 59 complete viral genome
TCGA-MY-A5BE-01A-21R-A26T-07	132342657	9685094	142027751	Yes	8	1997	206.193146	1	1834	189.36316	189.36316	NC_001526 » Human papillomavirus type 16
TCGA-C5-A2LY-01A-31R-A18M-07	115403904	8238399	123642303	Yes	6	2125	257.938467	2	2105	255.51081	189.235797	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9UH-01A-11R-A42T-07	108073460	4100497	112173957	Yes	5	860	209.730673	2	823	200.707377	187.294369	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-DR-A0ZL-01A-11R-A10U-07	162367597	24866508	187234105	Yes	14	5471	220.014808	3	5340	214.746678	185.711641	NC_001526 » Human papillomavirus type 16
TCGA-C5-A7CM-01A-11R-A33Z-07	127053759	8437368	135491127	Yes	8	1811	214.640395	2	1786	211.677386	176.476835	NC_001357 » Human papillomavirus - 18
TCGA-EA-A3Y4-01A-51R-A24H-07	196503717	6003638	202507355	Yes	9	1278	212.87093	2	1219	203.043555	173.394865	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-VS-A8EG-01A-11R-A36F-07	180384862	7805303	188190165	Yes	10	1897	243.039893	4	1890	242.143067	167.706494	NC_001526 » Human papillomavirus type 16
TCGA-EK-A3GJ-01A-21R-A213-07	175318540	11730585	187049125	Yes	12	2558	218.062439	2	2427	206.895052	162.055004	D90400 » Human papillomavirus type 58 complete genome
TCGA-C5-A1M8-01A-21R-A13Y-07	221446458	7756963	229203421	Yes	8	1539	198.402389	3	1518	195.695145	157.020215	NC_001526 » Human papillomavirus type 16
TCGA-ZJ-AAXU-01A-11R-A42T-07	214779103	5671490	220450593	Yes	5	1016	179.141637	2	1007	177.554752	155.691009	NC_001526 » Human papillomavirus type 16
TCGA-C5-A7UE-01A-11R-A33Z-07	118300424	10138082	128438506	Yes	6	1778	175.378341	2	1745	172.123287	152.494328	NC_001526 » Human papillomavirus type 16
TCGA-IR-A3LH-01A-21R-A213-07	190835879	9585111	200420990	Yes	9	1987	207.300676	2	1967	205.214108	150.650316	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-EK-A2PL-01A-11R-A18M-07	106338339	6791432	113129771	Yes	8	1974	290.660349	3	1962	288.893417	150.189238	M62849 » Human papillomavirus ORFs (HPV-39)
TCGA-EK-A2RO-01A-11R-A18M-07	129357098	5760956	135118054	Yes	9	925	160.563628	1	858	148.933614	148.933614	NC_001526 » Human papillomavirus type 16
TCGA-EX-A8YF-01A-11R-A37O-07	135341397	9864621	145206018	Yes	10	1857	188.248489	2	1809	183.382616	141.921317	NC_001357 » Human papillomavirus - 18
TCGA-BI-A0VR-01A-11R-A10U-07	175984156	28930625	204914781	Yes	14	4737	163.736522	2	4452	153.885372	140.439413	NC_001526 » Human papillomavirus type 16
TCGA-DG-A2KM-01A-11R-A180-07	210643129	9112359	219755488	Yes	8	1419	155.722573	2	1399	153.527752	134.76203	NC_001526 » Human papillomavirus type 16
TCGA-BI-A0VS-01A-11R-A10U-07	221754124	30977248	252731372	Yes	21	4936	159.342759	2	4792	154.694181	134.29211	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1ME-01A-11R-A13Y-07	122154298	11703576	133857874	Yes	14	1891	161.574549	2	1822	155.678914	133.20715	NC_001357 » Human papillomavirus - 18
TCGA-C5-A1BF-01B-11R-A13Y-07	177553931	7491679	185045610	Yes	6	1153	153.904085	3	1145	152.836234	127.341281	NC_001357 » Human papillomavirus - 18
TCGA-RA-A741-01A-11R-A33Z-07	104903999	13724681	118628680	Yes	9	1995	145.358569	2	1963	143.027004	125.321674	NC_001526 » Human papillomavirus type 16
TCGA-EK-A2RE-01A-11R-A18M-07	187611121	10821906	198433027	Yes	9	1502	138.792555	2	1489	137.591289	109.777335	NC_001526 » Human papillomavirus type 16
TCGA-EX-A1H6-01B-11R-A22U-07	182628042	11790202	194418244	Yes	11	1322	112.127001	1	1211	102.712405	102.712405	NC_001526 » Human papillomavirus type 16

TCGA-UC-A7PG-01A-11R-A42S-07	140198530	9063363	149261893	Yes	9	990	109.230976	2	948	104.596936	93.563504	NC_001526 » Human papillomavirus type 16
TCGA-EA-A3HR-01A-11R-A213-07	214074455	8346111	222420566	Yes	10	851	101.963657	1	775	92.85762	92.85762	NC_001587 » Human papillomavirus type 34
TCGA-UC-A7PD-01A-11R-A352-07	135115385	16236662	151352047	Yes	9	1691	104.147022	2	1657	102.052996	87.086866	NC_001526 » Human papillomavirus type 16
TCGA-IR-A3LC-01A-11R-A213-07	180551447	18192817	198744264	Yes	8	2917	160.338005	2	2873	157.919469	86.792496	NC_001526 » Human papillomavirus type 16
TCGA-DS-A0VL-01A-21R-A10U-07	165667301	19874272	185541573	Yes	11	3189	160.458706	2	3146	158.295106	86.795632	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1BK-01B-11R-A13Y-07	164923670	18616060	183539730	Yes	10	1948	104.640832	2	1900	102.062413	84.980388	NC_001526 » Human papillomavirus type 16
TCGA-UC-A7PG-06A-11R-A42S-07	150356407	14561775	164918182	Yes	7	1389	95.386723	2	1371	94.15061	83.918341	NC_001526 » Human papillomavirus type 16
TCGA-EA-A78R-01A-11R-A32P-07	112819875	7771944	120591819	Yes	9	636	81.832809	1	609	78.358774	78.358774	X74481 » Human papillomavirus type 52 genomic DNA
TCGA-C5-A1M6-01A-11R-A13Y-07	119282731	18406755	137689486	Yes	13	1934	105.070123	2	1800	97.790186	77.199919	NC_001357 » Human papillomavirus - 18
TCGA-FU-A3EO-01A-11R-A213-07	123736857	15949975	139686832	Yes	7	1269	79.561252	1	1227	76.92802	76.92802	NC_001526 » Human papillomavirus type 16
TCGA-EA-A50E-01A-21R-A26T-07	120493206	11757449	132250655	Yes	8	961	81.735415	1	829	70.508492	70.508492	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1MF-01A-11R-A13Y-07	154359388	17982449	172341837	Yes	13	1340	74.517103	1	1214	67.510271	67.510271	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-C5-A1BJ-01A-11R-A13Y-07	273528255	27074313	300602568	Yes	10	3028	111.840327	2	2999	110.7692	60.093861	NC_001526 » Human papillomavirus type 16
TCGA-C5-A1BM-01A-11R-A13Y-07	185889680	20159843	206049523	Yes	14	1619	80.308167	2	1419	70.387453	59.722687	NC_001357 » Human papillomavirus - 18
TCGA-C5-A7X5-01A-11R-A36F-07	154346700	5258163	159604863	Yes	15	645	122.66641	3	557	105.930531	54.391619	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-JW-A5VI-01A-11R-A28H-07	138202281	4435004	142637285	Yes	7	334	75.309967	2	324	73.055177	53.663988	NC_001593 » Human papillomavirus type 53
TCGA-VS-A8Q8-01A-11R-A37O-07	113586085	8598163	122184248	Yes	5	449	52.220457	1	433	50.359594	50.359594	NC_001526 » Human papillomavirus type 16
TCGA-DG-A2KJ-01A-11R-A32Y-07	72792235	5812677	78604912	Yes	6	347	59.697108	1	294	50.579105	50.579105	NC_001357 » Human papillomavirus - 18
TCGA-VS-A950-01A-11R-A42T-07	109262513	3120238	112382751	Yes	13	396	126.913396	3	319	102.235791	49.996186	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-EA-A44S-01A-12R-A26T-07	138101110	13070336	151171446	Yes	8	718	54.933552	1	632	48.353768	48.353768	NC_001526 » Human papillomavirus type 16
TCGA-DS-A1O9-01A-11R-A14Y-07	237218614	11660398	248879012	Yes	14	476	40.821932	1	432	37.048478	37.048478	NC_001587 » Human papillomavirus type 34
TCGA-C5-A7UI-01A-11R-A36F-07	101625249	31200653	132825902	Yes	14	1213	38.877393	1	1007	32.274966	32.274966	M12732 » Human papillomavirus type 33 complete genome
TCGA-ZX-AA5X-01A-11R-A42T-07	74267128	3106134	77373262	Yes	9	224	72.11537	2	141	45.394049	28.009094	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-Q1-A5R3-01A-11R-A28H-07	201863457	16316652	218180109	Yes	7	747	45.78145	2	668	40.939771	23.53424	NC_001357 » Human papillomavirus - 18
TCGA-JW-A852-01A-11R-A352-07	102227106	18576449	120803555	Yes	7	423	22.770768	1	400	21.532641	21.532641	NC_001526 » Human papillomavirus type 16
TCGA-VS-A9V0-01A-11R-A42T-07	52655226	1192819	53848045	Yes	3	44	36.887407	2	43	36.049057	20.958754	M12732 » Human papillomavirus type 33 complete genome
TCGA-IR-A3LA-01A-11R-A22U-07	170083246	6283647	176366893	Yes	9	108	17.187469	1	96	15.277752	15.277752	NC_001526 » Human papillomavirus type 16
TCGA-VS-A8EK-01A-12R-A37O-07	138990889	6164069	145154958	Yes	9	192	31.148259	1	89	14.438515	14.438515	J04353 » Human papillomavirus type 31 (HPV-31) complete genome
TCGA-FU-A3EO-11A-13R-A213-07	166691399	4740854	171432253	Yes	3	66	13.921541	1	64	13.499677	13.499677	NC_001526 » Human papillomavirus type 16
TCGA-C5-A8YT-01A-11R-A37O-07	138369094	5714932	144084026	Yes	3	84	14.698337	1	77	13.473476	13.473476	X74479 » Human papillomavirus type 45 genomic DNA
TCGA-VS-A8QH-01A-11R-A37O-07	121332119	8833464	130165583	Yes	5	98	11.094176	1	90	10.188529	10.188529	NC_001526 » Human papillomavirus type 16

Table 2. List of genes involved in significant pathways related to immune response and viral integration in CESC samples obtained through GSEA.

CESC	
GO_BP Pathways	Genes
Cellular defence response	FOSL1, FAIM3, IL1RL2, NCF2, HLA-G, GNLY, LYST, C5AR1, ADORA2B, MICB, KLRC2, CCL5, CLEC5A, GAGE8, TCIRG1, KLRC3, KLRC4, PRF1, LBP, CD5L, CD160, TRAT1, KIR2DL4, MNDA, CD19, NCR1, TYROBP, CCR6, ITK, GAGE1, LSP1, CCR5, LGALS3BP, BECN1, KIR3DL2, CXCL9, CX3CR1, CD300C, MICA, KLRG1, UMOD, LILRB2, ADORA2A, VEZF1, ZNF148, NCF1, LY96, SPN, NCR2, CCR3, CCR2, CXCR2, IL4, CCR9, ITGB1, PAGE1, DCDC2
Defence response	FOSL1, CCR7, HCP5, CCL22, ALOX5AP, NFKB1, NMI, HLA-B, FAIM3, GPR68, BCL10, MX1, TPSAB1, PYDC1, LTB4R, ANXA1, IL10RB, F11R, LY75, FOXN1, IL12B, IL1RL2, SELE, NCF2, HLA-G, HRH1, GNLY, S100A9, IL1RAP, FOS, IL1A, RNASE6, IRAK2, APOBEC3G, IL32, LYST, BLNK, PTAFR, MX2, IFNK, IL20, C5AR1, CHRNA7, ADORA2B, CXCL11, TLR3, MICB, RSAD2, KLRC2, CCL13, CXCL1, CCL5, CD40, CXCL10, PSG8, PTPRCAP, IL18RAP, APOBEC3F, CLEC5A, MST1R, GAGE8, NLRP3, CCR4, S100A12, CCL26, AIF1, NOS2, IL13, NFX1, CXCL6, CD97, INHBA, CX3CL1, IL29, RAC1, TCIRG1, CD48, WFDC12, SFTPD, KLRC3, KLRC4, CRTAM, PSG3, CFHR1, APOL3, PRF1, LBP, ALOX15, CD5L, CD160, LILRA2, CEBPB, KCNN4, TRAT1, CCL23, CCL21, TNIP1, KIR2DL4, MNDA, S100A7, SP140, BNIP3L, CD19, NCR1, CYSLTR1, NFATC3, TYROBP, CEBPG, TNFRSF1A, PTPRC, TPST1, ANKRD1, MEFV, ADORA3, CCR6, ITK, IL12A, GAGE1, C5, LSP1, CCR5, CCL11, PARP4, HDAC5, CCL4, CD40LG, LILRA3, CCL3, LGALS3BP, GATA3, ELF3, WAS, CLEC1A, C3AR1, CD1D, BECN1, IL8, NFE2L1, NOD2, PGLYRP2, CAMP, KIR3DL2, CXCL9, S100A8, NLRC4, CX3CR1, STAB1, CD300C, TGFB1, MICA, VPS45, PRDX5, TLR7, AIMP1, KLRG1, CFP, UMOD, LYZ, LILRB2, CXCL2, PGLYRP3, ADORA2A, MBL2, FPR2, IL5, CCL3L3, CYBB, VEZF1, MGLL, ADORA1, CCL20, CCR1, IL9, PLA2G2E, IFNA2, DEFB127, DEFB118, LILRB3, AFAP1L2, ZNF148, NCF1, IL28RA, AZU1, GHRL, OR2H2, LALBA, CD84, LY96, BCL2, AOX1, SPN, AOAH, NCR2, CD83, DEFB1, TLR6, TNFAIP6, LILRA1, CCR3, CCR2, CYP4F11, TLR8, SOCS6, C2, NOD1, AHSG, CXCR2, IL4, NOX4, LILRB5, HP, CCR9, MLF2, MPO, XCR1, GHSR, PLA2G2D, IL17RB, HDAC7, PTX3, CLEC1B, ABCF1, P2RY11, STAB2, PLA2G7, CXCR4, EREG, SPACA3, PGLYRP4, KRT1, RELA, ORM1, DMBT1, CSF3R, SIGIRR, NFRKB, ITGB1, CCL24, PAGE1, TFF3, IL17C, PGLYRP1, COLEC12, TIAL1, APCS, TARBP2, CDO1, BNIP3, TACR1, APOA4, NFATC4, HDAC9, INHBB, CRP, S1PR3, CAMLG, AGER, KNG1, ORM2, RIPK2, AOC3, CHST2, DCDC2, CD81, SCG2, HDAC4, CADM1
Immune response	MR1, IL4R, MS4A2, CCL22, GPR183, IFI6, BCL10, PYDC1, LTB4R, CCR8, IL10RB, MS4A1, LY75, TRAF2, IL12B, BST2, TNFAIP1, IL7R, S1PR4, APOBEC3G, IL32, CNR2, BLNK, IK, PTAFR, CTLA4, TRIM22, IL15, IFNK, C5AR1, RSAD2, IKBKG, MBP, LAT2, CCL5, CIITA, CCRL1, RGS1, APOBEC3F, CD7, IL16, CCR4, FOXP3, SPINK5, TREM1, GZMA, TNFRSF14, SLA2, CCL2, DEFB4A, TREM2, CCL26, PAX5, NFIL3, CD96, PSMB10, CD86, CD97, CRHR1, CX3CL1, IL29, GBP2, SFTPD, LAT, CRTAM, CFHR1, CST7, IL6, FYB, CTSW, NCF4, DPP8, CD28, CEBPB, TRAT1, CCL23, SEMA7A, CCL21, IL6R, BNIP3L, NCR1, GPR65, CEBPG, PTPRC, SKAP1, TAPBP, TNFRSF4, IFITM3, CCR6, KIR2DL1, IL12A, FCGR2B, AIM2, GTPBP1, IKBKAP, IL6ST, CCR5, ARHGDIB, LY86, KIR2DL3,

	CCL4, VIPR1, CD40LG, IL2RA, EBI3, IRF8, CTSG, IL18BP, CD74, WAS, ZAP70, CD22, FCGR3B, CD1D, LCP2, IL10, POU2AF1, IL7, PTGER4, FCN1, ST6GAL1, CD274, CD79B, APLN, POU2F2, IL18, CD79A, FCGR3A, ANXA11, TGFB1, BCAR1, ETS1, IL2RG, GPI, TLR7, SECTM1, CD164, HAMP, LILRB2, CCBP2, MBL2, CCL27, HRH2, SP2, CTSS, IL27, IL2, FCGR1A, CXCL13, CCL19, CCL20, CCR1, DEFB127, DEFB118, FCN2, PRELID1, TRAF6, CCL18, IL28RA, IGSF6, CHST4, DPP4, CTSC, BST1, BCL2, IFITM2, CD83, FTH1, DEFB1, IL1R2, LAX1, TGFB2, MADCAM1, LTF, IL17B, UBE2N, CCR2, TLR8, C2, IL4, AQP9, NFAM1, CEACAM8, IL27RA, ZEB1, MALT1, FCGRT, MNX1, CXCL12, GEM, CCR9, OPRD1, FCAR, PDCD1, CMKLR1, CXCR4, XBP1, EREG, RFX1, IL17A, MAP3K7, CNIH, SEMA4D, KRT1, DMBT1, CCL25, CCL24, GPR44, ODZ1, COLEC12, VTN, APOA2, SEMA3C, TARBP2, RAG1, CHUK, TCF12, BNIP3, APOA4, SOCS5, APOA1, THY1, TCF7, C1QBP, OPRK1, YTHDF2, TNFSF13, FYN, MAP4K2, PRKRA, ATP6V0A2, CADM1
Response to virus	FOSL1, POLA1, CCL22, ISG20, BCL3, IFNGR1, APOBEC3G, IFI44, TRIM22, IFNK, IRF7, RSAD2, CCL5, APOBEC3F, TNF, IL29, CCL8, BNIP3L, PTPRC, CCL11, CCL4, FGR, IFNAR2, HNRNPUL1, IFNA4, IFNAR1, TLR7, BANF1, IFNGR2, CREBZF, HBXIP, CCL19, IFNA17, IFNA7, IL28RA, LILRB1, BCL2, DUOX2, IVNS1ABP, IFNW1, TLR8, CCDC130, CXCL12, CXCR4, SPACA3, ABCE1, TARBP2, BNIP3, C19orf2, PRKRA
Inflammatory response	CCR7, CCL22, ALOX5AP, NFKB1, NMI, GPR68, LTB4R, ANXA1, IL10RB, F11R, LY75, SELE, HRH1, S100A9, IL1RAP, FOS, IL1A, IRAK2, BLNK, PTAFR, IL20, CHRNA7, CXCL11, CCL13, CXCL1, CCL5, CD40, CXCL10, IL18RAP, NLRP3, CCR4, S100A12, CCL26, AIF1, IL13, NFX1, CXCL6, CD97, CX3CL1, RAC1, CFHR1, APOL3, LBP, ALOX15, CEBPB, CCL23, CCL21, NFATC3, TNFRSF1A, TPST1, MEFV, ADORA3, C5, CCR5, CCL11, PARP4, HDAC5, CCL4, CD40LG, CCL3, ELF3, C3AR1, IL8, NFE2L1, CXCL9, S100A8, TGFB1, VPS45, PRDX5, AIMP1, KLRG1, LYZ, CXCL2, ADORA2A, MBL2, FPR2, IL5, CCL3L3, CYBB, MGLL, ADORA1, CCL20, CCR1, IL9, PLA2G2E, IFNA2, AFAP1L2, GHRL, AOX1, AOAH, TNFAIP6, CCR3, CCR2, CYP4F11, C2, NOD1, AHSG, CXCR2, NOX4, XCR1, GHSR, PLA2G2D, HDAC7, PTX3, ABCF1, PLA2G7, CXCR4, KRT1, RELA, ORM1, SIGIRR, NFRKB, CCL24, IL17C, APCS, CDO1, TACR1, NFATC4, HDAC9, CRP, S1PR3, AGER, KNG1, ORM2, RIPK2, AOC3, CHST2, SCG2, HDAC4
Immune system process	MR1, IL4R, MS4A2, CCL22, CD47, CKLF, GPR183, LIG1, IFI6, BCL10, SAA1, PYDC1, LTB4R, CCR8, IL10RB, MS4A1, LY75, TRAF2, IL12B, ELF4, CDC42, BST2, ERAP2, TNFAIP1, IL7R, S1PR4, CTSE, APOBEC3G, IL32, CNR2, BLNK, IK, CD3D, PTAFR, CTLA4, TRIM22, IFI16, IL15, IFNK, C5AR1, RSAD2, IKBKG, MBP, LAT2, CCL5, CIITA, CCRL1, CD3E, MAFB, RGS1, APOBEC3F, CD7, IL16, CCR4, FOXP3, CLEC7A, SPINK5, TREM1, GZMA, TNFRSF14, SLA2, NOTCH2, CCL2, DEFB4A, TREM2, CCL26, CD2, PAX5, NFIL3, CD96, PSMB10, CD86, CD97, CRHR1, INHBA, JAG2, CX3CL1, IL29, GBP2, SFTPD, LRMP, LAT, CRTAM, LST1, CFHR1, CST7, IL6, FYB, CTSW, NCF4, DPP8, CD28, CEBPB, TRAT1, CCL23, SEMA7A, CCL21, IL6R, HELLS, NCK1, BNIP3L, NCR1, GPR65, CEBPG, LCK, SIRPG, PTPRC, NLRC3, SKAP1, TAPBP, TNFRSF4, IFITM3, MMP9, TAZ, CCR6, KIR2DL1, IL12A, LYN, FCGR2B, AIM2, GTPBP1, IKBKAP, IL6ST, CCR5, ARHGDIB, LY86, KIR2DL3, ITGB2, HDAC5, CCL4, VIPR1, CD40LG, IL2RA, EBI3, IRF8, CTSG, IL18BP, HCLS1, RPS19, CD74, WAS, ZAP70, CD22, FCGR3B, CD1D, LCP2, SOD1, IL10, POU2AF1, IL7, IL8, PTGER4, FCN1, ST6GAL1, TBX1, CD274, CD79B, APLN, POU2F2, MAP4K1, IL18, CD79A, FCGR3A, ANXA11, DOCK2, CSF1, TGFB1, BCAR1, ETS1, MLF1, IL2RG, GPI, SPI1,

	RASGRP4, DYRK3, TLR7, MAL, AIMP1, RUNX1, SECTM1, CD164, IL31RA, HAMP, LILRB2, CCBP2, MBL2, CCL27, HRH2, SP2, SCIN, CTSS, NCOA6, IL27, IL2, FCGR1A, CD4, CXCL13, MYH9, CCL19, CCL20, CCR1, PRG3, DEFB127, DEFB118, FCN2, PRELID1, FOXO3, TRAF6, CCL18, TM7SF4, SART1, IL28RA, AZU1, IGSF6, CHST4, DPP4, CTSC, BST1, MLL, BCL2, ICOSLG, CD276, ZBTB16, IFITM2, CD83, FTH1, SIT1, DEFB1, IL1R2, LAX1, TGFB2, MADCAM1, LTF, GLMN, IL17B, UBE2N, CCR2, TLR8, C2, CXCR2, IL4, AQP9, NFAM1, CEACAM8, IL27RA, ZEB1, MALT1, FCGRT, CD34, SNRK, MNX1, IL21, CXCL12, GEM, CCR9, ACVR1B, OPRD1, FCAR, TPD52, PDCD1, CD24, HDAC7, PREX1, INHA, CMKLR1, CXCR4, SYK, XBP1, EREG, RFX1, IL17A, RAB3D, SPACA3, ACVR2A, MAP3K7, ALAS2, CNIH, SEMA4D, KRT1, DMBT1, CCL25, NOTCH4, CCL24, NCK2, GPR44, PF4, PRL, ODZ1, ZNF675, COLEC12, VTN, APOA2, ACIN1, INS, SEMA3C, TARBP2, RAG1, CHUK, TCF12, AKT1, CARTPT, BNIP3, APOA4, HDAC9, NHEJ1, SOCS5, APOA1, LDB1, THY1, TCF7, C1QBP, CDK6, LIG3, OPRK1, YTHDF2, TNFSF13, KIRREL3, CALCA, FYN, TLR4, MAP4K2, PRKRA, MIA3, SCG2, HDAC4, ATP6V0A2, CADM1
Endosome transport	ADRB2, EEA1, RAB14, VPS4B, TINAGL1, LYST, MON2, DOPEY2, RAB35, ABCA1, M6PR, STX5, TOM1, DOPEY1, ANKRD27, VTI1A, ZFYVE16, SQSTM1, GOSR1, YKT6, RHOB, STX16, BET1L
JAK-STAT cascade	STAT4, NMI, SOCS3, STAT5A, STAT3, STAT1, IL20, IL22RA2, SOCS2, FGFR3, CCL2, STAT2, IL29, PIGU, PIAS1, SOCS1, IL12A, LYN, HCLS1, STAMBP, IFNAR2, IFNAR1, IL31RA, CCR2, SOCS6, CLCF1, STAT5B, F2R, HGS, F2, NF2
Response to other organism	FOSL1, POLA1, CCL22, ISG20, BCL10, BCL3, IFNGR1, APOBEC3G, IFI44, TRIM22, IFNK, IRF7, TLR3, RSAD2, SLC11A1, CCL5, APOBEC3F, S100A12, TNF, NOS2, IL29, WFDC12, CCL8, S100A7, BNIP3L, PTPRC, CHIT1, IL12A, CCL11, CCL4, FGR, IFNAR2, CD1D, IL10, NOD2, PGLYRP2, HNRNPUL1, CAMP, NLRC4, STAB1, IFNA4, IFNAR1, TLR7, CFP, PGLYRP3, BANF1, IFNGR2, CREBZF, HBXIP, CCL19, IFNA17, IFNA7, DEFB127, DEFB118, IL28RA, AZU1, LALBA, LILRB1, BCL2, SPN, DUOX2, TLR6, IVNS1ABP, IFNW1, TLR8, NOD1, CCDC130, CXCL12, CD24, STAB2, CXCR4, SPACA3, PGLYRP4, ABCE1, DMBT1, ITLN1, PGLYRP1, TARBP2, CHIA, BNIP3, C19orf2, PRKRA
T-cell activation	CD47, IL12B, ELF4, CD3D, CD3E, CD7, FOXP3, CLEC7A, SPINK5, SLA2, CD2, JAG2, SFTPD, LAT, CRTAM, CD28, NCK1, LCK, SIRPG, PTPRC, NLRC3, EBI3, ZAP70, CD1D, IL7, IL18, IL27, IL2, CD4, SART1, ICOSLG, CD276, SIT1, LAX1, GLMN, IL4, IL21, CD24, NCK2, INS, NHEJ1, SOCS5, THY1, CADM1
keratinocyte differ-entiation	TGM1, ANXA1, IVL, IL20, DSP, CSTA, SPRR1B, SPRR1A, TXNIP, SCEL, EVPL, LOR, TGM3, NME2, EREG
GO MF_Pathways	Genes
Interleukin receptor activity	IL4R, IL12RB2, IL15RA, IL10RB, IL7R, IL22RA2, IL9R, IL2RB, IL6R, IL12RB1, IL6ST, IL10RA, IL2RA, IL1R1, IL3RA, IL2RG, CSF2RB, IL5RA, CXCR2, IL13RA2
Interleukin binding	IL4R, IL12RB2, IL15RA, IL10RB, IL7R, IL22RA2, IL9R, IL2RB, IL6R, IL12RB1, IL12A, IL6ST, IL10RA, IL2RA, EBI3, IL18BP, IL1R1, IL3RA, IL2RG, CSF2RB, HAX1, IL5RA, CXCR2, A2M, IL13RA2
Cytokine binding	IL4R, PLP2, IL12RB2, IL15RA, IL10RB, IL7R, IFNGR1, IL22RA2, IL9R, TNFRSF14, TNFRSF25, ELANE, IL2RB, IL22RA1, IL6R, IL12RB1, TNFRSF1A, TNFRSF18, TNFRSF4, IL12A, IL6ST, IL10RA, IL2RA, EBI3, IL18BP, IL1R1,

	TNFRSF1B, CD74, IL3RA, IFNAR2, IL2RG, IFNAR1, CSF2RB, ACVR1, IFNGR2, HAX1, IL17F, IL5RA, CXCR2, ACVRL1, TGFBR1, GPR17, XCR1, CNTFR, CMKLR1, A2M, CRLF1, IL13RA2
Hematopoietin interferon classd200 domain cytokine receptor activity	IL4R, OSMR, IL12RB2, IL15RA, IL10RB, IL7R, IFNGR1, IL22RA2, IL9R, IL2RB, IL22RA1, IL6R, IL12RB1, IL6ST, IL10RA, IL2RA, IL1R1, IL3RA, IFNAR2, IL2RG, IFNAR1, CSF2RB, IL31RA, IFNGR2, GHR, IL5RA, CXCR2, CNTFR, GFRA1, GFRA2, LIFR, IL13RA2, EPOR
Cytokine activity	TYMP, CCL22, CKLF, OSM, IL12B, CXCL16, SIVA1, ERAP1, IFNK, IL20, CCL17, CXCL11, IL1RN, CCL13, CXCL1, CCL5, CXCL10, MUC4, CCL2, CCL26, TNF, CXCL6, INHBA, CX3CL1, IL29, CSF2, CCL8, PIK3R1, CCL23, CCL21, IFNA13, CSF3, GDF15, IL12A, CXCL3, C5, CCL11, CCL4, CD40LG, CCL3, IFNA14, TRIP6, IL7, IL8, CXCL9, GH1, CSF1, IFNA4, CCL16, CCL15, IL25, AIMP1, SECTM1, YARS, CTF1, CXCL2, CCL28, IFNA8, CCL27, IFNG, XCL2, IL19, IL5, CXCL14, IL27, IL2, BRE, CXCL13, CCL19, IFNA5, CCL20, IFNA16, IFNA17, IFNA7, IFNA2, ERBB2IP, CCL18, MIF, CNTF, CCL7, TGFB2, IFNA21, IL3, CCL1, IL17F, GLMN, IL17B, IL4, ERBB2, IL21, CXCL12, FGF10, FIGF, INHA, NAMPT, IL17A, CCL25, CXCL5, CCL24, PF4, PRL, IL17C, VEGFA, SPRED1, SPRED2, CDK5, INHBB, TNFRSF11B, XCL1, BMP4, SDCBP, SCG2
Chemokine activity	CCL22, CKLF, CXCL16, CCL17, CXCL11, CCL13, CXCL1, CCL5, CXCL10, CCL2, CCL26, CXCL6, CX3CL1, CCL8, CCL23, CCL21, CXCL3, C5, CCL11, CCL4, CCL3, IL8, CXCL9, CCL16, CCL15, CXCL2, CCL28, CCL27, XCL2, CXCL14, CXCL13, CCL19, CCL20, CCL18, CCL7, CCL1, CXCL12, CCL25, CXCL5, CCL24, PF4, XCL1
Chemokine receptor binding	CCL22, CKLF, CXCL16, CCL17, CXCL11, CCL13, CXCL1, CCL5, CXCL10, CCL2, CCL26, CXCL6, CX3CL1, CCL8, CCL23, CCL21, CXCL3, C5, CCL11, CCL4, CCL3, IL8, CXCL9, CCL16, CCL15, CXCL2, CCL28, CCL27, XCL2, CXCL14, CXCL13, CCL19, CCL20, CCL18, CCL7, CCL1, CCR2, CXCL12, CCL25, CXCL5, CCL24, PF4, XCL1
Kegg Pathways	Genes
DNA replication	PRIM1, POLE2, POLA1, PCNA, POLD3, MCM5, LIG1, RFC4, RFC5, MCM6, RPA3, POLA2, POLE4, RPA2, POLD4, FEN1, MCM4, POLE3, MCM2, POLD1, RFC2, RPA1, RFC1, RNASEH2B, MCM7, PRIM2, POLE, MCM3, DNA2, POLD2, RFC3, RNASEH2A, RNASEH2C, SSBP1, RNASEH1, RPA4
Mismatch repair	PCNA, POLD3, LIG1, RFC4, MSH6, RFC5, RPA3, RPA2, POLD4, POLD1, RFC2, EXO1, MSH3, RPA1, RFC1, MLH3, POLD2, MLH1, MSH2, RFC3, SSBP1, RPA4, PMS2
Cytosolic DNA sensing pathway	PYCARD, NFKB1, TMEM173, POLR3GL, NFKBIA, RIPK1, TBK1, NFKBIB, DDX58, IRF7, IKBKG, CCL5, CXCL10, IRF3, TREX1, ADAR, IL6, IL1B, POLR3D, ZBP1, POLR3K, IFNA13, CASP1, IFNB1, AIM2, CCL4, RIPK3, IFNA14, IFNA1, IL18, IL33, IFNA4, POLR1D, IFNA8, IFNA5, IKBKE, POLR3F, IFNA16, IFNA17, IFNA10, IFNA7, IFNA6, IFNA2, POLR3C, CCL4L2, POLR3B, POLR3G, IKBKB, IFNA21, RELA, CHUK, POLR3H, POLR1C, POLR3A, MAVS
Cytokine receptor interaction	CCR7, IL4R, FLT3LG, OSMR, TNFRSF11A, CCL22, IL18R1, CXCR5, IL12RB2, IL15RA, IL24, CCR8, IL10RB, OSM, CRLF2, IL12B, NGFR, CXCL16, EGFR, IL1RAP, TNFRSF6B, TNFRSF13B, CSF2RA, IL7R, RELT, IL1A, TNFSF10, IFNGR1, IL11, TNFRSF12A, IL15, IFNK, IL20, CCL17, CXCL11, IL22RA2, LTB, TNFRSF10A, CCL13, CD70, CXCL1, CCL5, CD40, IFNE, CXCL10, IL9R, IL18RAP, CCR4, CXCR6, TNFRSF14, IL28A, CCL2, CCL26,

	TNF, IL13, TNFRSF25, CXCL6, INHBA, FLT3, CX3CL1, IL29, CSF2, LTA, CD27, IL2RB, TNFRSF9, IL6, IL13RA1, IL1B, CCL8, IL22RA1, IL23R, CCL23, CCL21, IL6R, IFNA13, IL28B, IL12RB1, TNFSF9, CCR10, FAS, MET, CSF3, TNFRSF1A, IL20RB, PDGFA, TNFRSF18, TNFRSF4, CCR6, IL12A, IFNB1, CXCL3, TNFRSF10B, TSLP, IL6ST, CCR5, CCL11, CCL14, CCL4, IL10RA, CD40LG, IL2RA, CCL3, IL1R1, GH2, TNFRSF1B, CCL3L1, IL3RA, IFNAR2, TNFSF13B, IFNA14, IL10, IL7, IL8, VEGFC, AMH, IFNA1, CXCL9, IL18, FASLG, CX3CR1, GH1, CSF1, TGFB1, IL2RG, IFNA4, CCL16, CCL15, IL25, IFNAR1, CSF1R, TNFRSF17, CSF2RB, TNFRSF8, CTF1, CXCL2, ACVR1, BMP7, TNFSF8, EGF, CCL28, IFNA8, IFNGR2, CCL27, LEP, IFNG, CXCR3, XCL2, IL19, IL5, IL21R, CCL3L3, CXCL14, IL2, IL17RA, TNFSF14, TNFSF12, CXCL13, CCL19, IFNA5, CCL20, CCR1, IL9, IFNA16, IFNA17, IFNA10, IFNA7, IFNA6, IFNA2, LTBR, CCL4L2, KITLG, CCL18, PLEKHO2, IL28RA, PDGFB, TNFSF4, CNTF, TPO, TNFSF18, IL11RA, TNFRSF10C, TGFBR2, TNFRSF21, IL1R2, HGF, CCL7, TGFB2, IFNA21, IL3, CCL1, TNFSF15, EDAR, IL17B, GHR, CCR3, CCR2, IL20RA, FLT4, IFNW1, IL5RA, CXCR2, IL4, CLCF1, ACVRL1, FLT1, TGFBR1, CXCR1, IL21, KDR, CXCL12, CCR9, ACVR1B, BMPR2, TGFB3, EPO, INHBC, PPBP, XCR1, CNTFR, TNFRSF13C, IL17RB, MPL, PDGFRA, TNFRSF10D, IL22, AMHR2, FIGF, LEPR, CXCR4, GDF5, IL17A, ACVR2A, PRLR, PDGFRB, CCL25, CSF3R, KIT, CXCL5, BMPR1B, CCL24, VEGFB, INHBE, PF4, PRL, PF4V1, VEGFA, IL23A, TNFSF11, LIF, LIFR, IL26, INHBB, TNFRSF11B, BMP2, PDGFC, XCL1, TNFSF13, EPOR, EDA2R, EDA, TNFRSF19, BMPR1A, ACVR2B
Proteasome	PSMB9, PSMD7, PSMB2, PSMA6, PSMB8, PSMA5, PSMA7, PSME1, PSMD2, PSMA2, PSMD8, PSME2, PSMB7, PSMA4, PSMC1, PSMB10, PSMC4, PSMB1, PSMD11, PSMC5, PSMB6, PSMB3, PSME4, POMP, PSMA3, PSMB4, SHFM1, PSMD4, PSMC3, PSMC2, PSMC6, PSMA8, IFNG, PSMD14, PSMA1, PSME3, PSMF1, PSMD13, PSMD3, PSMB5, PSMB11, PSMD12, PSMD6, PSMD1
Natural killer cell mediated cytotoxi-city	HLA-C, HLA-A, ULBP2, NFATC1, HLA-B, HLA-E, PTK2B, HLA-G, KIR2DS4, CD247, KLRC1, TNFSF10, IFNGR1, ICAM1, PPP3CC, SH2D1B, NCR3, MICB, KLRC2, TNFRSF10A, PTPN6, SOS2, RAET1E, PIK3CA, CHP2, TNF, HCST, RAC1, CSF2, RAC2, CD48, RAET1L, KLRC3, LAT, ULBP1, PRF1, PIK3R1, KLRD1, KLRK1, IFNA13, KIR2DL4, SH2D1A, NCR1, FAS, NFATC3, TYROBP, LCK, PIK3CD, KIR2DL1, KIR3DL1, MAPK3, IFNB1, FCER1G, TNFRSF10B, VAV2, KIR2DL3, ITGB2, SH3BP2, PIK3CB, ZAP70, FCGR3B, ITGAL, IFNAR2, LCP2, IFNA14, ARAF, PPP3R1, CD244, ICAM2, CASP3, SHC3, KIR3DL2, IFNA1, FASLG, FCGR3A, VAV3, IFNA4, MICA, HRAS, IFNAR1, MAP2K1, SOS1, PIK3CG, CHP, IFNA8, NFATC2, IFNGR2, RAET1G, IFNG, PRKCB, NFAT5, IFNA5, IFNA16, IFNA17, IFNA10, IFNA7, IFNA6, IFNA2, GZMB, PIK3R5, PPP3R2, SHC1, BID, VAV1, TNFRSF10C, NCR2, SHC2, SHC4, IFNA21, GRB2, PRKCA, KRAS, MAP2K2, ULBP3, MAPK1, NRAS, TNFRSF10D, SYK, PLCG2, BRAF, PPP3CA, PIK3R3, PIK3R2, PTPN11, PAK1, RAF1, PLCG1, NFATC4, PRKCG, FYN, RAC3, PPP3CB
Primary immunode-ficiency	TAP1, TAP2, ICOS, ADA, TNFRSF13B, IL7R, UNG, BLNK, CD3D, AICDA, IKBKG, CD40, CIITA, CD3E, DCLRE1C, CD19, LCK, PTPRC, IGLL1, JAK3, CD40LG, RFXANK, ZAP70, CD79A, IL2RG, BTK, RFXAP, CD4, RAG2, CD8A, RFX5, TNFRSF13C, AIRE, RAG1, CD8B
Antigen processing and preservation	B2M, HLA-C, HLA-A, HLA-B, HLA-E, TAP1, TAP2, HLA-F, HLA-G, KIR2DS4, KLRC1, CTSB, PSME1, KLRC2, CIITA, PSME2, IFI30, LTA, KLRC3, KLRC4, HSPA2, KIR3DL3, HLA-DRB1, KLRD1, HLA-DQB1, IFNA13,

	KIR2DL4, HSPA6, HLA-DPB1, TAPBP, CTSL1, KIR2DL1, KIR3DL1, HLA-DRA, KIR2DL3, HLA-DRB5, HLA-DQA2, CD74, RFXANK, HLA-DMA, HSPA1A, IFNA14, HLA-DPA1, KIR3DL2, IFNA1, IFNA4, HSPA5, NFYB, HLA-DOB, IFNA8, RFXAP, CTSS, HSP90AA1, CD4, IFNA5, IFNA16, IFNA17, IFNA10, IFNA7, IFNA6, IFNA2, HSPA1L, HSPA1B, LGMN, HLA-DQA1, NFYC, PDIA3, CREB1, PSME3, IFNA21, CD8A, CALR, HLA-DOA, RFX5, HLA-DMB, HSPA8, CANX, CD8B, HSPA4, HSP90AB1, NFYA
GO_CC Pathways	Genes
Chromosome	STAG3, POLA1, DSN1, TINF2, PCNA, RFC4, RFC5, NPM2, SMC1A, CDT1, RPA3, SUMO3, JUN, XRCC4, ING2, RPA2, SMC2, RGS12, SMC4, TMPO, TIPIN, KIF22, CLIP1, SUV39H1, CENPC1, PSEN1, ZWILCH, MCM2, RFC2, ACD, HMGB2, TIMELESS, CENPA, CHMP1A, DMC1, PMF1, DCTN2, RAD51, SYCE2, PURA, RPA1, ERCC1, HELLS, OIP5, RFC1, RB1, CHEK1, BUB3, MCM7, HIST4H4, PAFAH1B1, UBE2I, ZWINT, HMGN1, FOXC1, APTX, MIS12, APC, ERCC4, MCM3, MAF, BIRC5, CENPE, JUND, JUNB, NOL6, HDAC8, BCL6, CENPF, NCAPD2, NDC80, CDCA5, TOP2A, ZW10, RAN, ZBED1, RCC1, TERF2, BUB1B, RFC3, ATRX, NSL1, TUBG1, PIF1, INCENP, POLG2, H1FNT, ZNF238, CBX5, MAD2L1, H2AFY, ZNF330, TNKS, BUB1, TOP1, RPA4, ZMIZ2, TTN, CLASP1, SYCE1, KLHDC3, HMGB1, LIG4, SMARCA5, PURB, SUGT1, TNP1, UPF1, TERF2IP, MKI67IP, LRPPRC, PAM, CBX1, NUFIP1, PSEN2, SUV39H2, H2AFY2, SMARCE1, MYCN, TOP2B, DNMT3A, REPIN1
Chromosomeperi-centric region	DSN1, SMC1A, SUMO3, KIF22, CLIP1, CENPC1, PSEN1, ZWILCH, CENPA, PMF1, DCTN2, HELLS, BUB3, PAFAH1B1, ZWINT, MIS12, APC, BIRC5, CENPE, CENPF, NDC80, ZW10, BUB1B, NSL1, INCENP, MAD2L1, ZNF330, BUB1, CLASP1, SUGT1, PSEN2
Condensed Chro-mosome	STAG3, DSN1, SMC1A, XRCC4, SMC2, RGS12, SMC4, SUV39H1, HMGB2, CHMP1A, DMC1, PMF1, RAD51, SYCE2, CHEK1, UBE2I, MIS12, CENPE, NOL6, CENPF, NCAPD2, RCC1, NSL1, TUBG1, TTN, SYCE1, HMGB1, LIG4, SMARCA5, MKI67IP, LRPPRC, PAM
Condensed Nuclear Chromosome	STAG3, SMC1A, RGS12, SUV39H1, CHMP1A, DMC1, RAD51, SYCE2, CHEK1, UBE2I, NOL6, RCC1, TUBG1, TTN, SYCE1, MKI67IP, LRPPRC
Nuclear Chromo-some Part	STAG3, POLA1, NPM2, RPA3, RPA2, TIPIN, ACD, TIMELESS, SYCE2, PURA, RPA1, ERCC1, MCM7, UBE2I, FOXC1, ERCC4, MCM3, RCC1, ATRX, PIF1, H1FNT, CBX5, H2AFY, RPA4, ZMIZ2, SYCE1, KLHDC3, PURB, PAM, CBX1, NUFIP1, H2AFY2, REPIN1
Chromosomal part	STAG3, POLA1, DSN1, TINF2, PCNA, RFC4, RFC5, NPM2, SMC1A, CDT1, RPA3, SUMO3, RPA2, SMC2, SMC4, TMPO, TIPIN, KIF22, CLIP1, CENPC1, PSEN1, ZWILCH, MCM2, RFC2, ACD, TIMELESS, CENPA, PMF1, DCTN2, SYCE2, PURA, RPA1, ERCC1, HELLS, OIP5, RFC1, RB1, BUB3, MCM7, HIST4H4, PAFAH1B1, UBE2I, ZWINT, HMGN1, FOXC1, APTX, MIS12, APC, ERCC4, MCM3, MAF, BIRC5, CENPE, JUND, JUNB, BCL6, CENPF, NDC80, CDCA5, ZW10, RAN, RCC1, TERF2, BUB1B, RFC3, ATRX, NSL1, PIF1, INCENP, H1FNT, CBX5, MAD2L1, H2AFY, ZNF330, TNKS, BUB1, RPA4, ZMIZ2, CLASP1, SYCE1, KLHDC3, PURB, SUGT1, TNP1, UPF1, PAM, CBX1, NUFIP1, PSEN2, SUV39H2, H2AFY2, MYCN, TOP2B, DNMT3A, REPIN1
Proteasome com-plex	PSMD7, PSME1, PSMD2, PSMD8, PSME2, PSMC4, PSMD10, PSMD11, PSMC5, ADRM1, KIAA0368, SHFM1, PSMD4, PSMC3, PSMC2, PSMC6, PSMD14, PSME3, PSMD13, PSMD5, PSMD3, PSMD12, PSMD1

Table 3. *LIHC resume table of viral infection.*

LIHC Samples	Nº Reads	Human Reads	Non-Human Reads	Viral Infection	Virus Found	Total Human Viral Reads	Total ppm of Human Viral Reads	Human Virus with ppm>10	Sum of all Reads with ppm>10	Sum of all ppm bigger than 10	Bigger ppm found	Virus with bigger ppm
TCGA-DD-AAEK-01A-11R-A41C-07	70305983	2549987	72855970	Yes	1	10367	4065.510922	1	10367	4065.510922	4065.510922	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADP-01A-11R-A39D-07	120792108	5308131	126100239	Yes	2	13923	2622.95712	1	13916	2621.638388	2621.638388	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADV-01A-11R-A39D-07	175339586	6929501	182269087	Yes	6	10724	1547.586184	2	10718	1546.72032	1531.134782	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACA-01A-11R-A41C-07	151429927	6551245	157981172	Yes	2	9488	1448.274335	1	9486	1447.96905	1447.96905	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADU-01A-11R-A41C-07	118532667	4460172	122992839	Yes	5	5895	1321.697907	1	5858	1313.402263	1313.402263	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-A116-11A-12R-A26B-07	164021083	5908704	169929787	Yes	1	7010	1186.385373	1	7010	1186.385373	1186.385373	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAD0-01A-11R-A41C-07	116619732	2917337	119537069	Yes	1	3387	1160.990314	1	3387	1160.990314	1160.990314	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACA-02B-11R-A41C-07	147739426	4580425	152319851	Yes	1	5263	1149.020015	1	5263	1149.020015	1149.020015	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-A1EA-01A-11R-A131-07	119368196	2714982	122083178	Yes	1	3107	1144.390644	1	3107	1144.390644	1144.390644	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADL-01A-11R-A41C-07	145068722	5905214	150973936	Yes	2	6584	1114.946893	1	6582	1114.608209	1114.608209	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAE2-01A-11R-A41C-07	154007360	5528320	159535680	Yes	3	6152	1112.815467	1	6147	1111.911033	1111.911033	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACA-02A-11R-A41C-07	137698888	9739493	147438381	Yes	2	9983	1025.00202	1	9982	1024.899345	1024.899345	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACS-01A-11R-A41C-07	81221104	1966642	83187746	Yes	1	1987	1010.351655	1	1987	1010.351655	1010.351655	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADD-01A-11R-A41C-07	103924450	3328756	107253206	Yes	1	3210	964.32421	1	3210	964.32421	964.32421	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-G3-A25U-01A-11R-A16W-07	144672990	4185121	148858111	Yes	1	3716	887.907423	1	3716	887.907423	887.907423	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACN-01A-11R-A41C-07	112002050	5588980	117591030	Yes	2	4904	877.440965	1	4903	877.262041	877.262041	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACD-01A-11R-A41C-07	115179921	4355831	119535752	Yes	1	3683	845.533263	1	3683	845.533263	845.533263	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADG-01A-11R-A41C-07	126862944	4273798	131136742	Yes	1	3528	825.495262	1	3528	825.495262	825.495262	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADF-01A-11R-A41C-07	109953620	4444063	114397683	Yes	1	3243	729.73763	1	3243	729.73763	729.73763	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACV-01A-11R-A41C-07	93954755	7653309	101608064	Yes	2	5502	718.904724	1	5501	718.774062	718.774062	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACE-01A-11R-A41C-07	99542909	3144155	102687064	Yes	1	2135	679.03777	1	2135	679.03777	679.03777	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-G3-A3CH-11A-11R-A22L-07	130874869	4910062	135784931	Yes	2	3224	656.610854	1	3213	654.370556	654.370556	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A3MB-01A-11R-A213-07	172470251	5675462	178145713	Yes	3	3611	636.247763	1	3609	635.895369	635.895369	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAC9-01A-11R-A41C-07	105757426	5027831	110785257	Yes	2	3072	610.999057	1	3071	610.800164	610.800164	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAVP-01A-11R-A41C-07	71806971	1917621	73724592	Yes	1	1154	601.787319	1	1154	601.787319	601.787319	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-ED-A7XP-01A-11R-A352-07	128149090	4219283	132368373	Yes	5	2503	593.228754	1	2497	591.806712	591.806712	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADM-01A-11R-A41C-07	121165330	3054419	124219749	Yes	1	1780	582.762221	1	1780	582.762221	582.762221	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A7IK-01A-12R-A33R-07	141663929	4556810	146220739	Yes	2	2639	579.13321	1	2638	578.913758	578.913758	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-2Y-A9H4-01A-11R-A38B-07	89704718	3192920	92897638	Yes	2	1838	575.648623	1	1836	575.022237	575.022237	NC_003977 » Hepatitis B virus (strain ayw) genome

TCGA-DD-A116-01A-11R-A131-07	129114971	4657198	133772169	Yes	2	2606	559.563927	1	2594	556.98727	556.98727	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACT-01A-11R-A41C-07	70785343	2699939	73485282	Yes	3	1494	553.345835	1	1492	552.605077	552.605077	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-ED-A8O6-01A-11R-A36F-07	136720850	6909177	143630027	Yes	1	3799	549.848412	1	3799	549.848412	549.848412	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A123-01A-11R-A131-07	126196774	6224414	132421188	Yes	1	3373	541.898402	1	3373	541.898402	541.898402	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACW-01A-11R-A41C-07	111630959	6796395	118427354	Yes	1	3490	513.507529	1	3490	513.507529	513.507529	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-UB-A7MC-01A-11R-A33R-07	174291370	7284614	181575984	Yes	4	3547	486.916672	1	3532	484.857537	484.857537	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADC-01A-11R-A41C-07	79662146	2997752	82659898	Yes	1	1449	483.362199	1	1449	483.362199	483.362199	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-QA-A7B7-01A-11R-A32O-07	179421506	5568424	184989930	Yes	3	2637	473.563076	1	2635	473.203908	473.203908	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACQ-01A-11R-A41C-07	145940758	5518663	151459421	Yes	2	2508	454.4579	1	2507	454.276697	454.276697	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-G3-A3CK-01A-11R-A213-07	182140953	6627479	188768432	Yes	3	2970	448.1342	1	2968	447.832426	447.832426	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-BW-A5NP-01A-11R-A27V-07	143882626	5297759	149180385	Yes	1	2342	442.073715	1	2342	442.073715	442.073715	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-BC-A10W-11A-11R-A131-07	139476978	5500511	144977489	Yes	3	2408	437.777508	1	2405	437.232104	437.232104	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-A119-11A-11R-A131-07	139917728	5972110	145889838	Yes	3	2546	426.314987	1	2543	425.812652	425.812652	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAE3-01A-11R-A41C-07	92166571	7024566	99191137	Yes	1	2974	423.371351	1	2974	423.371351	423.371351	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAD6-01A-11R-A41C-07	72689256	4665642	77354898	Yes	1	1811	388.156657	1	1811	388.156657	388.156657	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-G3-A25Z-01A-11R-A16W-07	176263715	4673251	180936966	Yes	2	1811	387.524659	1	1810	387.310675	387.310675	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-G3-AAV1-01A-11R-A38B-07	139722555	3251606	142974161	Yes	1	1247	383.502798	1	1247	383.502798	383.502798	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-BC-A10W-01A-11R-A131-07	158686279	5629997	164316276	Yes	3	2164	384.369654	1	2160	383.659174	383.659174	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A5UD-01A-11R-A28V-07	113890691	4179505	118070196	Yes	2	1597	382.102665	1	1594	381.384877	381.384877	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-5264-01A-01R-A131-07	182762654	8180852	190943506	Yes	6	3109	380.033766	1	3066	374.77759	374.77759	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-G3-AAV0-01A-11R-A37K-07	110875134	7983570	118858704	Yes	1	2729	341.827027	1	2729	341.827027	341.827027	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-A11A-01A-11R-A131-07	153811676	4723364	158535040	Yes	4	1636	346.363313	1	1605	339.800193	339.800193	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADR-01A-11R-A41C-07	105546520	7949768	113496288	Yes	3	2620	329.569367	1	2617	329.191997	329.191997	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACP-01A-11R-A41C-07	105739285	5482233	111221518	Yes	2	1804	329.062993	1	1794	327.238919	327.238919	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-5C-AAPD-01A-21R-A39D-07	107403930	4365579	111769509	Yes	2	1430	327.562506	1	1428	327.104377	327.104377	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-A1EH-11A-11R-A131-07	153296943	6346057	159643000	Yes	2	2046	322.40492	1	2044	322.089764	322.089764	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAD5-01A-11R-A41C-07	138243163	6656224	144899387	Yes	2	2121	318.649132	1	2119	318.348661	318.348661	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACK-01A-11R-A41C-07	297195465	13045146	310240611	Yes	4	3933	301.491451	1	3927	301.03151	301.03151	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-RC-A7SB-01A-21R-A352-07	160585269	7531328	168116597	Yes	2	2250	298.752093	1	2249	298.619314	298.619314	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-A1EL-11A-11R-A155-07	189341448	8625236	197966684	Yes	2	2567	297.615045	1	2566	297.499106	297.499106	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAVW-01A-11R-A41C-07	107402980	2252658	109655638	Yes	1	661	293.431138	1	661	293.431138	293.431138	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADK-01A-11R-A41C-07	217476210	5667329	223143539	Yes	1	1605	283.202193	1	1605	283.202193	283.202193	NC_003977 » Hepatitis B virus (strain ayw) genome

TCGA-DD-A1EI-01A-11R-A131-07	179005382	9698836	188704218	Yes	3	2686	276.940449	1	2682	276.528029	276.528029	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A9FS-01A-11R-A37K-07	132932504	8042963	140975467	Yes	5	2264	281.488302	1	2189	272.163381	272.163381	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A9FU-01A-11R-A37K-07	149731807	10595946	160327753	Yes	7	2903	273.972707	1	2873	271.141435	271.141435	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A8HT-01A-11R-A36F-07	120177349	4112849	124290198	Yes	1	1093	265.752523	1	1093	265.752523	265.752523	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACC-01A-11R-A41C-07	101844100	2743024	104587124	Yes	4	748	272.691745	1	726	264.671399	264.671399	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A9FW-01A-11R-A37K-07	157970876	14571315	172542191	Yes	5	3846	263.943234	1	3828	262.70793	262.70793	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A5UC-01A-11R-A28V-07	95246463	2488787	97735250	Yes	2	654	262.778614	1	653	262.376812	262.376812	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-RC-A7S9-01A-11R-A33R-07	149288714	7633147	156921861	Yes	3	1896	248.390343	1	1892	247.866313	247.866313	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAVS-01A-11R-A41C-07	143619299	4731031	148350330	Yes	2	1130	238.848572	1	1129	238.637202	238.637202	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-A1EI-11A-11R-A131-07	170306508	9821702	180128210	Yes	2	2282	232.342622	1	2281	232.240807	232.240807	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-A1EL-01A-11R-A155-07	202763897	8184656	210948553	Yes	4	1780	217.480124	1	1768	216.013966	216.013966	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-ZS-A9CF-02A-11R-A38B-07	125447430	3326319	128773749	Yes	3	808	242.911159	2	807	242.610526	206.534611	NC_001401 » Adeno-associated virus - 2
TCGA-CC-A7IL-01A-11R-A33R-07	170102848	6186113	176288961	Yes	2	1261	203.843673	1	1260	203.682021	203.682021	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A5UE-01A-11R-A28V-07	118614217	8007365	126621582	Yes	3	1615	201.68932	1	1609	200.94001	200.94001	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAVV-01A-11R-A41C-07	113773940	6912205	120686145	Yes	2	1363	197.187438	1	1361	196.898095	196.898095	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAVQ-01A-11R-A41C-07	143933966	3594225	147528191	Yes	3	697	193.922194	1	695	193.365746	193.365746	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAEE-01A-11R-A41C-07	146909251	8683866	155593117	Yes	4	1666	191.850035	1	1660	191.159099	191.159099	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-UB-A7ME-01A-11R-A33J-07	137082425	7847284	144929709	Yes	1	1458	185.796767	1	1458	185.796767	185.796767	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-ZS-A9CF-01A-11R-A38B-07	114291705	3040523	117332228	Yes	3	644	211.805667	2	642	211.147885	180.232151	NC_001401 » Adeno-associated virus - 2
TCGA-CC-A7IG-01A-11R-A33J-07	134679947	6832383	141512330	Yes	3	1237	181.04957	1	1224	179.146866	179.146866	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-BW-A5NO-01A-11R-A27V-07	69371126	2174331	71545457	Yes	5	402	184.884455	1	390	179.365515	179.365515	NC_004102 » Hepatitis C virus genotype 1
TCGA-G3-A3CJ-01A-11R-A213-07	140239378	4857483	145096861	Yes	3	884	181.987256	1	843	173.54667	173.54667	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-ED-A7XO-01A-11R-A352-07	100062337	11379880	111442217	Yes	1	1820	159.931388	1	1820	159.931388	159.931388	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A7II-01A-11R-A33J-07	149995731	7544334	157540065	Yes	2	1187	157.336619	1	1185	157.071519	157.071519	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAD1-01A-11R-A41C-07	103205530	3386345	106591875	Yes	1	480	141.745747	1	480	141.745747	141.745747	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A8HS-01A-11R-A36F-07	124175914	3639743	127815657	Yes	2	509	139.845039	1	496	136.273358	136.273358	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A1HT-01A-11R-A131-07	211315313	10525266	221840579	Yes	5	1357	128.927858	1	1332	126.552621	126.552621	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-A119-01A-11R-A131-07	139208308	6106845	145315153	Yes	3	747	122.321756	1	745	121.994254	121.994254	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-G3-AAV7-01A-11R-A38B-07	115166415	3241311	118407726	Yes	1	388	119.70465	1	388	119.70465	119.70465	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-K7-A5RG-01A-11R-A28V-07	121204308	5740945	126945253	Yes	3	1332	232.01755	2	1331	231.843363	118.447398	NC_007605 » Human herpesvirus 4 complete wild type genome
TCGA-DD-AAE6-01A-11R-A41C-07	119527821	8064701	127592522	Yes	2	949	117.673302	1	948	117.549305	117.549305	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADO-01A-11R-A41C-07	133437448	4380264	137817712	Yes	1	496	113.235184	1	496	113.235184	113.235184	NC_003977 » Hepatitis B virus (strain ayw) genome

TCGA-CC-A3MC-01A-11R-A22L-07	130149809	4923556	135073365	Yes	2	525	106.630248	1	523	106.224038	106.224038	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAD3-01A-11R-A41C-07	128928168	3631565	132559733	Yes	1	365	100.507632	1	365	100.507632	100.507632	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAE4-01A-11R-A41C-07	95963487	13501057	109464544	Yes	4	1324	98.066397	1	1316	97.47385	97.47385	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAE0-01A-11R-A41C-07	54294469	2853088	57147557	Yes	1	273	95.685797	1	273	95.685797	95.685797	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-5258-01A-01R-A131-07	167273080	7867384	175140464	Yes	1	731	92.915256	1	731	92.915256	92.915256	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-EP-A12J-11A-11R-A131-07	85325945	1934322	87260267	Yes	6	174	89.953999	1	164	84.784229	84.784229	NC_004102 » Hepatitis C virus genotype 1
TCGA-RG-A7D4-01A-12R-A33R-07	112303322	5193397	117496719	Yes	6	442	85.108071	1	422	81.257027	81.257027	NC_004102 » Hepatitis C virus genotype 1
TCGA-DD-A3A3-01A-11R-A22L-07	113103848	5428176	118532024	Yes	3	447	82.348104	1	445	81.979656	81.979656	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-5262-01A-01R-A131-07	115574935	5524615	121099550	Yes	3	426	77.109445	1	423	76.566421	76.566421	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAEB-01A-11R-A41C-07	94766527	17555085	112321612	Yes	3	1257	71.603185	1	1254	71.432294	71.432294	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A7IE-01A-21R-A38B-07	113936414	4980508	118916922	Yes	3	355	71.27787	1	353	70.876304	70.876304	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-5263-01A-01R-A131-07	159417127	5376674	164793801	Yes	4	422	78.487183	2	408	75.883343	63.794085	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADA-01A-11R-A41C-07	129667935	6949364	136617299	Yes	1	408	58.710409	1	408	58.710409	58.710409	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A8HU-01A-11R-A36F-07	114800545	3800558	118601103	Yes	1	214	56.307521	1	214	56.307521	56.307521	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAD2-01A-11R-A41C-07	94558874	3250259	97809133	Yes	1	176	54.149531	1	176	54.149531	54.149531	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-2Y-A9HA-01A-11R-A39D-07	166224345	5133144	171357489	Yes	8	299	58.248902	1	280	54.547466	54.547466	NC_004102 » Hepatitis C virus genotype 1
TCGA-DD-AADN-01A-11R-A41C-07	74846180	19537816	94383996	Yes	5	1051	53.793116	1	1030	52.718277	52.718277	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-BC-A10Y-01A-11R-A131-07	157767649	5886385	163654034	Yes	3	435	73.899346	2	433	73.559579	46.038443	NC_001401 » Adeno-associated virus - 2
TCGA-DD-AAED-01A-12R-A41C-07	92674598	16941737	109616335	Yes	2	732	43.206904	1	726	42.852749	42.852749	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACO-01A-11R-A41C-07	77523906	28323575	105847481	Yes	3	998	35.235666	1	994	35.09444	35.09444	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A3M9-01A-11R-A213-07	143077895	7906795	150984690	Yes	3	271	34.274317	1	268	33.894897	33.894897	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AAVU-01A-11R-A41C-07	121889622	3564414	125454036	Yes	3	120	33.666123	1	117	32.82447	32.82447	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-EP-A12J-01A-11R-A131-07	121179511	3472578	124652089	Yes	7	118	33.980517	1	108	31.100813	31.100813	NC_004102 » Hepatitis C virus genotype 1
TCGA-DD-AACU-01A-11R-A41C-07	114400017	4614271	119014288	Yes	1	126	27.306589	1	126	27.306589	27.306589	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACF-01A-11R-A41C-07	127573548	6464344	134037892	Yes	2	164	25.369937	1	163	25.215242	25.215242	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACB-01A-11R-A41C-07	79164411	5337481	84501892	Yes	3	176	32.974356	1	133	24.918121	24.918121	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADY-01A-11R-A41C-07	74991128	3081874	78073002	Yes	1	68	22.064497	1	68	22.064497	22.064497	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-A7IJ-01A-11R-A33R-07	136224206	5770539	141994745	Yes	3	112	19.408932	1	104	18.02258	18.02258	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACH-01A-11R-A41C-07	162207670	5238087	167445757	Yes	2	90	17.181845	1	89	16.990936	16.990936	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-BC-A10Q-11A-11R-A131-07	137951153	7936190	145887343	Yes	1	127	16.002641	1	127	16.002641	16.002641	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AADW-01A-11R-A39D-07	162750012	7899238	170649250	Yes	3	111	14.051988	1	108	13.672205	13.672205	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-RC-A6M3-01A-11R-A32O-07	140797149	4498398	145295547	Yes	4	63	14.004985	1	58	12.893479	12.893479	NC_003977 » Hepatitis B virus (strain ayw) genome

TCGA-DD-A3A3-11A-11R-A22L-07	102687232	3558051	106245283	Yes	1	44	12.366321	1	44	12.366321	12.366321	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-ED-A66Y-01A-11R-A311-07	127544015	3803926	131347941	Yes	2	48	12.618542	1	44	11.566997	11.566997	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CC-5259-01A-31R-A213-07	84953221	18295332	103248553	Yes	1	209	11.423679	1	209	11.423679	11.423679	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-G3-A25T-01A-11R-A16W-07	156872955	4101211	160974166	Yes	1	43	10.484708	1	43	10.484708	10.484708	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-DD-AACJ-01A-11R-A41C-07	190817423	17321852	208139275	Yes	4	208	12.007955	1	177	10.218307	10.218307	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-5R-AA1C-01A-11R-A41C-07	56201161	2440167	58641328	Yes	2	27	11.064816	1	26	10.655008	10.655008	NC_003977 » Hepatitis B virus (strain ayw) genome

Table 4. *Count and ratio (number of mutations found divided by the number of sample) of somatic mutations present in 193 samples in LIHC.*

	Samples	Infection (%)	Frame Shift Del.	Frame Shift Del. (ratio)	Frame Shift Ins.	Frame Shift Ins. (ratio)	In Frame Del.	In Frame Del. (ratio)
Inf.	45	23.32	668	14.84	234	5.20	172	3.82
Non-Inf.	148	76.68	1206	8.15	624	4.22	402	2.72
Total	193	100.00	1874	9.71	858	4.45	574	2.97

	Samples	In Frame Ins.	In Frame Ins. (ratio)	Missense Mutation	Missense Mutation (ratio)	Nonsense Mutation	Nonsense Mutation (ratio)	Nonstop Mutation	Nonstop Mutation (ratio)
Inf.	45	45	1.00	4780	106.22	302	6.71	14	0.31
Non-Inf.	148	131	0.89	16006	108.15	954	6.45	36	0.24
Total	193	176	0.91	20786	107.70	1256	6.51	50	0.26

	Samples	RNA	RNA (ratio)	Silent	Silent (ratio)	Splice Site	Splice Site (ratio)	Translation Start Site	Translation Start Site (ratio)
Inf.	45	6714	149.20	1819	40.42	366	8.13	39	0.87
Non-Inf.	148	11415	77.13	6107	41.26	1190	8.04	132	0.89
Total	193	18129	93.93	7926	41.07	1556	8.06	171	0.89

Table 5. *List of genes bearing somatic mutations in immune-related pathways significantly over-represented in the infected group in LIHC.*

Pathway	Genes
Phosphatidylinositol phospholipase C activity	CASR, PLCB4, PLCH1, PLCL1, PLCG1, CHRM3, PLCB2, PLCE1, PLCB3, EDNRA, PLCL2, CCR5, PLCD3, CCR1, BDKRB2, CHRM1, C3AR1, PLCB1, PLCD1, PLCG2, CCL5
Inflammatory mediator regulation of TRP channels	PRKCH, MAP2K3, MAPK9, PIK3CB, PRKCQ, CAMK2A, PRKACA, ASIC4, ADCY2, GNAS, ITPR3, PLCB4, TRPA1, PLA2G4C, PIK3R2, PIK3CG, MAPK10, PLA2G4A, PIK3CA, ITPR2, PLCG1, PTGER2, PRKCG, CYP2J2, HTR2B, PLCB2, PIK3R5, PRKACB, TRPM8, PIK3R1, PLCB3, ITPR1, ADCY8, GNAQ, PLA2G4D, ADCY9, F2RL1, ADCY1, PRKCB, BDKRB2, PLA2G4F, JMJD7-PLA2G4B, CALML6, PTGER4, PIK3CD, PPP1CA, ADCY5, ADCY6, P2RY2, PLCB1, PLA2G6, MAPK11, PPP1CC, PLA2G4E, IL1RAP, HRH1, TRPV1, PLCG2, NTRK1, ASIC3, ASIC5
Leukocyte aggregation	MAD1L1, CYP26B1, ITGAL, FYN, CD6, SLC11A1, CD44, RC3H2, ATG5, CASP8, AP3D1, PRKCQ, PRKCZ, BCL3, RORA, JMJD6, CDC42, STK10, TBX21, GLI2, DLG1, PAG1, PAK3, APBB1IP, PSEN1, TCF7, CYLD, BAX, SIRPG, LAG3, ICAM1, CD209, IL12RB1, MYH9, SLA2, BMP7, BMX, ELF4, CSK, RIPK2, CLPTM1, RASAL3, TGFB1, JAK3, PIK3CG, LFNG, EPHB6, GLI3, DOCK8, GATA3, PPP3CB, FOXN1, IFNG, PTPN6, CD83, SRF, ITK, IL4, BCL6, CD86, HHLA2, FOXP1, ZAP70, IL1RL2, IL18R1, SOS1, SLAMF1, PLA2G2D, MYB, SLC46A2, NR4A3, GPAM, IFNA8, TNFSF11, HSPH1, EGR1, CD80, PIK3CA, LAX1, PREX1, RUNX2, BCL11B, EIF2AK4, RIPK3, AP3B1, SPINK5, PTPN22, RSAD2, IL6ST, DOCK2, MAP3K7, IL6, HLX, IL36B, RPS6, IRF4, PDE5A, LEF1, PPP3CA, PDPK1, ERBB2, ITPKB, SOX13, HSPD1, PIK3R1, TNFRSF21, GNRH1, IFNA16, CDKN2A, GSN, ZEB1, SCGB1A1, PAK1, ADAM8, ADAM17, CAMK4, ZFP36L2, CD8A, CXADR, ADK, BATF, KIT, NCK1, CD1D, PLA2G2F, PKNOX1, AIRE, CD3G, C1ORF177, IL23R, NLRP3, SLAMF6, FZD5, EOMES, SPTA1, CTLA4, IL15, F2RL1, CASP3, RICTOR, SHH, ATP7A, RAG1, STAT6, NOD2, IGF2, NFKBID, CTNNB1, IL12A, LGALS9, PRELID1, HAS2, SOCS5, CHD7, PTGER4, PIK3CD, CTPS1, PRNP, RASGRP1, THEMIS, SP3, CD7, LIG4, LEP, TRAF6, PTPN2, SART1, YES1, PAWR, PTPN11, EGR3, PAK2, FUT7, CLEC4G, SATB1, LCK, JAG2, BTLA, IFNA10, HLA-DRB1, CD55, CD47, SRC, SPN, PDCD1LG2, KIF13B, CARD11, MTOR, BTNL2, HLA-G, PSMB10, LAT, HLA-DPB1, IFNA14, IFNA13, IFNA17, HLA-DQA2, HLA-DMB, PRKDC, INS, CCL5
Somatic diversification of immune receptors	ERCC1, TBX21, MLH1, TCF7, MSH2, CD40, NBN, IL27RA, TGFB1, EXOSC3, SUPT6H, IFNG, AICDA, IL4, BCL6, FOXP1, MSH6, BCL11B, LEF1, HSPD1, VPRBP, ATM, DCLRE1C, BATF, PAXIP1, RNF168, RAG1, STAT6, EXO1, LIG4, CLCF1, PRKDC

Table 6. *HNSC resume table of viral infection.*

HNSC Sample	Nº Reads	Human Reads	Non Human Reads	Viral Infection	Virus Found	Total Human Viral Reads	Total ppm of Human Viral Reads	Human Virus with ppm>10	Sum of all Reads with ppm>10	Sum of all ppm bigger than 10	Bigger ppm found	Virus with bigger ppm
TCGA-BA-5558-01A-01R-1514-07	178296042	4432752	182728794	Yes	2	44417	10020.186105	2	44417	10020.186105	8446.220316	NC_001806 » Human herpesvirus 1 strain 17
TCGA-BB-7866-01A-11R-2232-07	142680545	3536686	146217231	Yes	6	9936	2809.409714	2	9909	2801.775447	2789.899923	NC_001526 » Human papillomavirus type 16
TCGA-HD-A634-01A-11R-A28V-07	123216743	3988746	127205489	Yes	3	9506	2383.205148	2	9500	2381.700916	2251.584834	NC_001526 » Human papillomavirus type 16
TCGA-CR-7385-01A-11R-2016-07	84968417	2786891	87755308	Yes	3	5310	1905.349007	2	5309	1904.990184	1849.013829	NC_001526 » Human papillomavirus type 16
TCGA-BA-A4IH-01A-11R-A266-07	149753372	5110362	154863734	Yes	5	8142	1593.233513	1	8138	1592.450789	1592.450789	NC_001526 » Human papillomavirus type 16
TCGA-BA-5153-01A-01R-1436-07	113588728	3229575	116818303	Yes	4	4894	1515.369671	2	4890	1514.131117	1447.868528	NC_001526 » Human papillomavirus type 16
TCGA-QK-A6IF-01A-11R-A31N-07	93210195	2195620	95405815	Yes	1	2884	1313.524198	1	2884	1313.524198	1313.524198	NC_001526 » Human papillomavirus type 16
TCGA-CN-5374-01A-01R-1436-07	132205310	2676214	134881524	Yes	3	3567	1332.853053	2	3566	1332.479391	1295.113171	NC_001526 » Human papillomavirus type 16
TCGA-P3-A5QE-01A-11R-A28V-07	141754948	4883442	146638390	Yes	7	6939	1420.924012	2	6930	1419.081050	1294.578701	NC_001526 » Human papillomavirus type 16
TCGA-P3-A5QF-01A-11R-A28V-07	129671800	3994934	133666734	Yes	4	5235	1310.409634	2	5231	1309.408366	1290.634589	NC_001526 » Human papillomavirus type 16
TCGA-BB-4223-01A-01R-1436-07	122395774	2941968	125337742	Yes	4	3914	1330.401963	2	3909	1328.702420	1269.218428	NC_001526 » Human papillomavirus type 16
TCGA-DQ-7596-01A-11R-2232-07	186902827	4127016	191029843	Yes	4	5356	1297.789978	2	5336	1292.943861	1195.294615	NC_001526 » Human papillomavirus type 16
TCGA-BA-A4IG-01A-11R-A266-07	137500248	4862032	142362280	Yes	5	6322	1300.279389	2	6311	1298.016961	1191.682819	NC_001526 » Human papillomavirus type 16
TCGA-CR-5250-01A-01R-1514-07	140959816	4054266	145014082	Yes	5	4866	1200.217252	2	4849	1196.024138	1133.620734	NC_001526 » Human papillomavirus type 16
TCGA-CV-6433-01A-11R-1686-07	130216642	3571443	133788085	Yes	3	4351	1218.275079	2	4350	1217.995080	1132.035427	NC_001526 » Human papillomavirus type 16
TCGA-CR-5249-01A-01R-1514-07	165138012	4483339	169621351	Yes	6	5132	1144.682568	2	5121	1142.229040	1079.329491	NC_001526 » Human papillomavirus type 16
TCGA-DQ-7590-01A-11R-2232-07	132065535	3582294	135647829	Yes	4	3708	1035.090923	2	3694	1031.182812	993.776614	NC_001526 » Human papillomavirus type 16
TCGA-CN-A499-01A-11R-A24H-07	135674179	4462691	140136870	Yes	2	4269	956.597712	1	4264	955.477312	955.477312	NC_001526 » Human papillomavirus type 16
TCGA-QK-A6V9-01A-11R-A34R-07	147402470	6258677	153661147	Yes	2	6246	997.974492	2	6246	997.974492	937.258785	NC_001526 » Human papillomavirus type 16
TCGA-BA-4077-01B-01R-1436-07	122867820	3359582	126227402	Yes	2	3160	940.593205	1	3135	933.151803	933.151803	NC_001526 » Human papillomavirus type 16
TCGA-CR-5243-01A-01R-1514-07	167713985	4865922	172579907	Yes	2	4771	980.492495	2	4771	980.492495	909.796746	NC_001526 » Human papillomavirus type 16
TCGA-CR-5248-01A-01R-2016-07	126888267	4113519	131001786	Yes	4	3872	941.286524	2	3840	933.507296	907.738605	NC_001526 » Human papillomavirus type 16
TCGA-BB-4228-01A-01R-1436-07	149357498	4842695	154200193	Yes	4	4641	958.350670	2	4637	957.524684	888.554823	NC_001526 » Human papillomavirus type 16
TCGA-BA-5559-01A-01R-1514-07	207623967	6146672	213770639	Yes	2	5027	817.840939	2	5027	817.840939	799.945076	NC_001526 » Human papillomavirus type 16
TCGA-P3-A5Q5-01A-11R-A28V-07	147803156	4521324	152324480	Yes	5	3527	780.081232	1	3511	776.542446	776.542446	NC_001526 » Human papillomavirus type 16
TCGA-CN-4741-01A-01R-1436-07	184712981	7415768	192128749	Yes	7	5885	793.579304	2	5851	788.994478	767.553678	NC_001526 » Human papillomavirus type 16
TCGA-RS-A6TP-01A-12R-A34R-07	175517410	5885287	181402697	Yes	4	4512	766.657598	1	4497	764.108870	764.108870	NC_001526 » Human papillomavirus type 16
TCGA-CR-7404-01A-11R-2132-07	186133140	6742338	192875478	Yes	4	5139	762.198514	1	5134	761.456931	761.456931	NC_001526 » Human papillomavirus type 16
TCGA-DQ-7594-01A-11R-2232-07	168595744	6026470	174622214	Yes	4	4448	738.077183	1	4443	737.247510	737.247510	NC_001526 » Human papillomavirus type 16

TCGA-HD-7754-01A-11R-2081-07	202359466	7242726	209602192	Yes	7	5404	746.127908	2	5368	741.157404	728.178865	NC_001526 » Human papillomavirus type 16
TCGA-CV-5443-01A-01R-1514-07	107803780	5153354	112957134	Yes	2	3918	760.281556	2	3918	760.281556	727.681428	NC_001526 » Human papillomavirus type 16
TCGA-CV-6961-01A-21R-1915-07	219188595	7056839	226245434	Yes	5	5113	724.545367	1	5076	719.302226	719.302226	NC_001526 » Human papillomavirus type 16
TCGA-T2-A6X0-01A-11R-A34R-07	183179180	5981510	189160690	Yes	1	3973	664.213551	1	3973	664.213551	664.213551	NC_001526 » Human papillomavirus type 16
TCGA-CR-6472-01A-11R-1873-07	129132136	4515880	133648016	Yes	6	2993	662.772263	1	2982	660.336413	660.336413	NC_001526 » Human papillomavirus type 16
TCGA-CN-A6UY-01A-12R-A34R-07	118815651	5039976	123855627	Yes	4	3571	708.535121	4	3571	708.535121	642.066550	NC_001526 » Human papillomavirus type 16
TCGA-CR-6480-01A-11R-1873-07	155416151	7408274	162824425	Yes	3	4978	671.951389	2	4977	671.816405	641.579942	NC_001526 » Human papillomavirus type 16
TCGA-DQ-7593-01A-11R-2232-07	165708392	5994003	171702395	Yes	6	4154	693.026013	4	4131	689.188844	636.135818	NC_001526 » Human papillomavirus type 16
TCGA-CR-6487-01A-11R-1873-07	189054842	7061599	196116441	Yes	6	4161	589.243315	2	4155	588.393649	577.064770	NC_001526 » Human papillomavirus type 16
TCGA-MZ-A6I9-01A-11R-A31N-07	134307185	7303926	141611111	Yes	3	4287	586.944610	2	4282	586.260047	569.419789	NC_001526 » Human papillomavirus type 16
TCGA-BB-7861-01A-11R-2232-07	174434590	5014836	179449426	Yes	6	2836	565.521982	1	2762	550.765768	550.765768	NC_001526 » Human papillomavirus type 16
TCGA-CN-A49C-01A-11R-A24H-07	124974532	3740336	128714868	Yes	3	1997	533.909253	1	1993	532.839830	532.839830	NC_001526 » Human papillomavirus type 16
TCGA-CR-6470-01A-11R-1873-07	158273594	8076161	166349755	Yes	4	4405	545.432415	2	4396	544.318024	525.125737	NC_001526 » Human papillomavirus type 16
TCGA-CR-7369-01A-11R-2132-07	172407521	3424838	175832359	Yes	1	1718	501.629566	1	1718	501.629566	501.629566	NC_001526 » Human papillomavirus type 16
TCGA-UP-A6WW-01A-12R-A34R-07	180911163	6970629	187881792	Yes	3	3417	490.199665	2	3416	490.056206	456.917159	NC_001526 » Human papillomavirus type 16
TCGA-HL-7533-01A-11R-2232-07	135132943	5936385	141069328	Yes	4	2665	448.926410	1	2653	446.904977	446.904977	NC_001526 » Human papillomavirus type 16
TCGA-CR-7368-01A-11R-2132-07	198925741	4297086	203222827	Yes	3	1909	444.254549	1	1906	443.556401	443.556401	NC_001526 » Human papillomavirus type 16
TCGA-KU-A6H7-01A-11R-A31N-07	147223139	4832235	152055374	Yes	2	2158	446.584241	2	2158	446.584241	435.823175	NC_001526 » Human papillomavirus type 16
TCGA-BA-7269-01A-11R-2016-07	179436339	5524223	184960562	Yes	3	4422	800.474564	2	4421	800.293543	405.486889	NC_009334 » Human herpesvirus 4
TCGA-CV-5442-01A-01R-1514-07	135098526	3699540	138798066	Yes	3	1482	400.590345	1	1475	398.698217	398.698217	NC_001526 » Human papillomavirus type 16
TCGA-CQ-A4CE-01A-11R-A266-07	147271524	4139643	151411167	Yes	4	1630	393.753761	1	1593	384.815792	384.815792	NC_006273 » Human herpesvirus 5 strain Merlin
TCGA-CV-7406-01A-11R-2081-07	178604642	3196035	181800677	Yes	2	1179	368.894584	1	1178	368.581696	368.581696	NC_001526 » Human papillomavirus type 16
TCGA-CQ-5323-01A-01R-1686-07	105148902	2957031	108105933	Yes	3	1007	340.544282	1	984	332.766210	332.766210	NC_001526 » Human papillomavirus type 16
TCGA-H7-A76A-01A-51R-A34R-07	111590609	8668029	120258638	Yes	2	3044	351.175567	2	3044	351.175567	330.178868	NC_001526 » Human papillomavirus type 16
TCGA-CN-A6V7-01A-12R-A34R-07	168327652	7820521	176148173	Yes	4	2554	326.576708	1	2550	326.065233	326.065233	NC_001526 » Human papillomavirus type 16
TCGA-P3-A6T6-01A-11R-A34R-07	76553500	10714136	87267636	Yes	3	3338	311.551020	1	3335	311.271016	311.271016	NC_001526 » Human papillomavirus type 16
TCGA-CR-6473-01A-11R-1873-07	122259215	4298884	126558099	Yes	3	1240	288.446955	1	1201	279.374833	279.374833	NC_001526 » Human papillomavirus type 16
TCGA-CR-6481-01A-11R-1873-07	192691897	12676274	205368171	Yes	8	3737	294.802717	2	3712	292.830527	278.236333	NC_001526 » Human papillomavirus type 16
TCGA-CN-A63T-01A-11R-A28V-07	138983576	3282137	142265713	Yes	2	985	300.109350	2	985	300.109350	274.516268	NC_001806 » Human herpesvirus 1 strain 17
TCGA-CN-A6V6-01A-12R-A34R-07	99608301	10868175	110476476	Yes	6	2909	267.662235	1	2893	266.190046	266.190046	NC_001526 » Human papillomavirus type 16
TCGA-CV-5971-01A-11R-1686-07	180411398	6129940	186541338	Yes	3	1527	249.105211	1	1508	246.005671	246.005671	NC_001526 » Human papillomavirus type 16
TCGA-CR-6482-01A-11R-1873-07	154178104	16003375	170181479	Yes	9	3708	231.701126	1	3529	220.515985	220.515985	NC_001526 » Human papillomavirus type 16
TCGA-TN-A7HL-01A-11R-A34R-07	159013405	10179024	169192429	Yes	5	2320	227.919690	2	2313	227.232002	213.674710	NC_001526 » Human papillomavirus type 16

TCGA-BB-A6UM-01A-12R-A34R-07	150961103	6887740	157848843	Yes	4	1474	214.003433	1	1454	211.099722	211.099722	NC_001526 » Human papillomavirus type 16
TCGA-KU-A6H7-06A-21R-A31N-07	155248111	7054785	162302896	Yes	2	1130	160.174974	2	1130	160.174974	149.118648	NC_001526 » Human papillomavirus type 16
TCGA-P3-A6SW-01A-11R-A34R-07	110477818	7115906	117593724	Yes	2	752	105.678743	1	751	105.538213	105.538213	NC_001526 » Human papillomavirus type 16
TCGA-CQ-A4CH-01A-11R-A266-07	145105130	3788177	148893307	Yes	2	357	94.240580	1	356	93.976601	93.976601	NC_006273 » Human herpesvirus 5 strain Merlin
TCGA-BB-7864-01A-11R-2232-07	168951816	3593200	172545016	Yes	5	196	54.547479	1	145	40.354002	40.354002	NC_001593 » Human papillomavirus type 53
TCGA-UF-A7J9-01A-12R-A34R-07	163504031	8265279	171769310	Yes	3	318	38.474200	1	316	38.232224	38.232224	NC_006273 » Human herpesvirus 5 strain Merlin
TCGA-CN-A6V1-01A-12R-A34R-07	122071735	8337650	130409385	Yes	3	322	38.619994	1	317	38.020305	38.020305	NC_001526 » Human papillomavirus type 16
TCGA-QK-A8ZB-01A-11R-A39I-07	228156611	7006637	235163248	Yes	1	241	34.395959	1	241	34.395959	34.395959	NC_006273 » Human herpesvirus 5 strain Merlin
TCGA-CV-7438-11A-01R-2132-07	169290651	6212166	175502817	Yes	4	205	32.999762	1	200	32.194890	32.194890	NC_009333 » Human herpesvirus 8
TCGA-QK-A64Z-01A-11R-A30B-07	149647966	4756730	154404696	Yes	3	174	36.579752	1	142	29.852441	29.852441	NC_006273 » Human herpesvirus 5 strain Merlin
TCGA-BB-7862-01A-21R-2232-07	178399393	3925905	182325298	Yes	3	128	32.603947	1	104	26.490707	26.490707	NC_001526 » Human papillomavirus type 16
TCGA-CV-5970-01A-11R-1686-07	145450222	8979265	154429487	Yes	5	291	32.407998	1	230	25.614569	25.614569	NC_006273 » Human herpesvirus 5 strain Merlin
TCGA-DQ-7589-01A-11R-2232-07	219008217	4034911	223043128	Yes	3	108	26.766390	1	87	21.561814	21.561814	NC_001526 » Human papillomavirus type 16
TCGA-CN-A642-01A-12R-A30B-07	139134715	4008903	143143618	Yes	2	87	21.701698	1	86	21.452253	21.452253	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-H7-A6C5-11A-11R-A30B-07	129937690	5482433	135420123	Yes	3	120	21.888093	1	105	19.152081	19.152081	NC_001526 » Human papillomavirus type 16
TCGA-CV-5436-01A-01R-1514-07	134988322	4884792	139873114	Yes	2	101	20.676418	1	95	19.448116	19.448116	NC_006273 » Human herpesvirus 5 strain Merlin
TCGA-F7-A61S-01A-11R-A28V-07	96279012	6341880	102620892	Yes	3	220	34.690029	2	218	34.374665	17.345014	NC_009334 » Human herpesvirus 4
TCGA-D6-6826-01A-11R-1915-07	219244422	20802869	240047291	Yes	2	352	16.920743	1	339	16.295829	16.295829	NC_006273 » Human herpesvirus 5 strain Merlin
TCGA-BA-A4IF-01A-11R-A266-07	134538615	3548393	138087008	Yes	3	72	20.290875	1	58	16.345427	16.345427	NC_001526 » Human papillomavirus type 16
TCGA-CN-A49B-01A-31R-A24H-07	135288392	4003670	139292062	Yes	4	66	16.484876	1	61	15.236021	15.236021	NC_001526 » Human papillomavirus type 16
TCGA-BA-6870-01A-11R-1873-07	219307265	9732023	229039288	Yes	2	163	16.748830	1	151	15.515787	15.515787	NC_006273 » Human herpesvirus 5 strain Merlin
TCGA-HD-A633-01A-11R-A28V-07	165013079	4791625	169804704	Yes	4	72	15.026217	1	68	14.191428	14.191428	NC_001526 » Human papillomavirus type 16
TCGA-CN-A641-01A-11R-A30B-07	168368761	4322442	172691203	Yes	3	59	13.649691	1	56	12.955639	12.955639	NC_003977 » Hepatitis B virus (strain ayw) genome
TCGA-CV-7091-11A-01R-2016-07	115894814	2685963	118580777	Yes	2	32	11.913790	1	31	11.541484	11.541484	NC_001806 » Human herpesvirus 1 strain 17
TCGA-CQ-5332-01A-01R-1686-07	122274030	2393307	124667337	Yes	1	28	11.699293	1	28	11.699293	11.699293	NC_006273 » Human herpesvirus 5 strain Merlin
TCGA-CN-A49A-01A-11R-A24H-07	150104592	4436461	154541053	Yes	2	50	11.270245	1	49	11.044840	11.044840	NC_001526 » Human papillomavirus type 16
TCGA-BA-A4II-01A-11R-A266-07	141864878	4660809	146525687	Yes	5	135	28.964927	2	107	22.957387	11.585971	NC_007605 » Human herpesvirus 4 complete wild type genome
TCGA-CN-4735-01A-01R-1436-07	171643158	3614630	175257788	Yes	3	40	11.066139	1	38	10.512833	10.512833	NC_001716 » Human herpesvirus 7

Table 7. *List of genes in significant pathways related to immune response and viral integration in host DNA in HNSC samples obtained through GSEA.*

HNSC	
GO_BP Pathways	Genes
Chromosome Organization and Biogenesis	STAG3, NUSAP1, SMC4, CHAF1B, CHAF1A, RBBP4, MSH2, EZH2, SUV39H2, CENPH, MSH3, HDAC6, ZWINT, RBM14, ESPL1, SMARCC2, HELLS, DDX11, PDS5B, HIRIP3, TOP2A, ARID1A, NSD1, EHMT1, PBRM1, RPS6KA5, CDC23, RFC1, NASP, SMC1A, SMARCC1, SIRT1, SATB1, TSSK6, DFFB, TERT, INO80, HDAC2, CREBBP, NDC80, REC8, BPTF, NAP1L4, ACTL6A, ACIN1, SYCP1, MTA2, HUWE1, MIS12, CENPE, NCAPH, PIF1, TAF6L, WHSC1L1, SYCP3, NPM2, KIF25, HDAC5, LATS1, HDAC3, HDAC1, SMARCA5, SAFB, NBN, NAP1L1, SET, PAPD7, HDAC8, KDM4A, TLK2, SUPT16H, SMARCE1, MRE11A, SIRT5, SUPT4H1, RSF1, PPARGC1A, HDAC11, H1FNT, HMGB1, CDCA5, SMG6, AIFM2, POT1, TEP1, PTGES3, MAP3K12, ERCC4, SMARCD1, SIRT4, RAD50, HDAC4, PRMT8, TERF2, HDAC10, HMGA1, DKC1, NAP1L3, CARM1, VCX, KAT2A, TINF2, SIRT2, HDAC7, ASF1A, KAT5, ACD, TERF2IP, TNP1, PRMT7, TLK1, UBE2N, PHB, ERCC1, HMGA2, PAM, BNIP3, ZW10, NAP1L2, CHMP1A
DNA dependent DNA replication	RPA2, CDK2, MCM3, HMGB2, MSH6, PRIM1, MSH2, POLA1, MSH3, MLH1, TP73, TSPYL2, RPA1, RFC4, RFC3, RFC1, RAD17, PMS1, EXO1, GMNN, CDT1, CCDC88A, MUTYH, TIPIN, ATR, PRIM2, POLG2, CDC6, PURA, CDC45, MSH5, PRKCG, REV3L, RPAIN, NBN, POLB, ABL1, RAD51, TFAM, RAD9A, WRNIP1, GLI2, GLI1, ENPP7, NF2, RPA4, S100A11, PMS2, GTPBP4, CDK2AP1, EREG
DNA replication	MCM2, RPA2, MCM5, CDK2, MCM3, HMGB2, MSH6, PRIM1, MSH2, POLA1, MSH3, MLH1, POLE, TP73, TSPYL2, RPA1, RFC4, REV1, POLE2, MCM7, RFC3, NT5M, RFC1, RAD17, NASP, POLD1, PMS1, EXO1, REPIN1, GMNN, CDT1, RNASEH2A, CCDC88A, POLE3, MUTYH, DBF4, TIPIN, ATR, DUT, SPHAR, PRIM2, NOL8, CDKN2D, POLG2, CDC6, PURA, CDC45, MSH5, KCTD13, PRKCG, REV3L, RPAIN, NBN, NAP1L1, POLB, SET, NUP98, ABL1, RAD51, RPA3, MCM3AP, ACHE, MRE11A, TFAM, KIN, RAD9A, WRNIP1, GLI2, KRT7, POT1, UPF1, GLI1, ENPP7, NF2, RAD50, TERF2, RPA4, IGF1, DKC1, PTMS, TINF2, EGF, TERF2IP, POLD2, S100A11, PMS2, TYMP, GTPBP4, CDK2AP1, IGHMBP2, POLD4, EREG, NAE1, TBRG1, RBMS1
DNA replication initiation	MCM3, POLA1, RAD17, CDT1, TIPIN, CDC6, PURA, CDC45, NBN, RAD9A, WRNIP1, RPA4
GO_CC Pathways	Genes
T-cell receptor signalling	MAP3K14, PIK3R3, NCK1, CD40LG, RAF1, PIK3CB, CD28, LCK, CD8B, CD3E, LAT, ZAP70, DLG1, ITK, IKBKB, CD3G, PTPRC, GRAP2, PDCD1, IL2, SOS1, CD247, PIK3CG, GSK3B, PAK3, CD4, CD3D, PTPN6, CD8A, LCP2, NFATC3, TEC, MAP3K8, MAP2K7, VAV1, MAPK13, PIK3R1, PIK3R5, NFATC2, GRB2, NFATC1, PAK2, VAV3, PDK1, NFKB1, CHP, PPP3R1, AKT3, PIK3CA, IFNG, IL4, PAK4, PPP3CC, RASGRP1, KRAS, ICOS, CTLA4, CDK4, NFKBIE, MAPK9, FYN, MALT1, NRAS, PAK6, CBLC, RHOA, CHP2, MAPK1, PRKCQ, PIK3R2, SOS2, NCK2, AKT2, IL5, NFKBIB, FOS, PAK7, MAPK14, CBLB, PLCG1, BCL10, CBL, CARD11, CHUK, MAPK11, MAP2K1, MAP2K2, JUN, PPP3R2, IKBKG, NFAT5, PPP3CA, PAK1, NFKBIA, PPP3CB, CDC42, IL10, MAP3K7, PIK3CD, VAV2, NFATC4, TNF, MAPK3, MAPK12, RELA, CSF2, HRAS, AKT1

B-cell receptor signaling	SYK, PIK3R3, CR2, RAF1, PIK3CB, PRKCB, INPP5D, BTK, IKBKB, CD22, SOS1, CD19, CD79A, PIK3CG, GSK3B, PTPN6, NFATC3, CD79B, RASGRP3, LYN, BLNK, VAV1, PIK3AP1, PIK3R1, PIK3R5, NFATC2, GRB2, NFATC1, CD72, VAV3, NFKB1, CHP, PPP3R1, AKT3, PIK3CA, FCGR2B, PPP3CC, KRAS, PLCG2, NFKBIE, MALT1, NRAS, RAC2, CHP2, RAC1, MAPK1, LILRB3, PIK3R2, SOS2, DAPP1, AKT2, NFKBIB, FOS, BCL10, CARD11, CHUK, RAC3, MAP2K1, MAP2K2, JUN, PPP3R2, IKBKG, NFAT5, PPP3CA, NFKBIA, PPP3CB, PIK3CD, VAV2, IFITM1, NFATC4, MAPK3, RELA, HRAS, AKT1, CD81
Chromosomal part	TIMELESS, TMPO, MCM2, STAG3, RPA2, MCM3, RFC5, SMC4, SMC2, SUV39H2, NSL1, POLA1, CBX5, PCNA, ZWINT, CENPF, TNKS, HELLS, DNMT3A, RPA1, RFC4, OIP5, ZWILCH, MCM7, BUB1B, RFC3, RFC1, CENPC1, SMC1A, DSN1, ZMIZ2, TOP2B, SYCE2, APC, RCC1, REPIN1, CDT1, HIST4H4, NDC80, RFC2, TIPIN, BCL6, MYCN, ATRX, PMF1, MIS12, CENPE, HMGN1, CENPA, PURA, PIF1, NPM2, ZNF330, H2AFY, MAD2L1, KIF22, BUB3, BUB1, SYCE1, CBX1, RB1, INCENP, CLASP1, JUNB, RPA3, KLHDC3, H2AFY2, H1FNT, CDCA5, SUMO3, UPF1, ERCC4, TERF2, RPA4, APTX, CLIP1, JUND, UBE2I, FOXC1, BIRC5, DCTN2, TINF2, PURB, PAFAH1B1, MAF, ACD, TNP1, PSEN1, NUFIP1, ERCC1, PAM, ZW10, RAN, PSEN2, SUGT1
Chromosome	TIMELESS, TMPO, MCM2, STAG3, RPA2, MCM3, RFC5, SMC4, HMGB2, SMC2, SUV39H2, NSL1, POLA1, CBX5, PCNA, ZWINT, CENPF, TNKS, HELLS, DNMT3A, TOP2A, RPA1, RFC4, DMC1, OIP5, ZWILCH, MCM7, BUB1B, ZNF238, RFC3, SUV39H1, RFC1, CENPC1, SMC1A, DSN1, ZMIZ2, TOP2B, SYCE2, APC, RCC1, REPIN1, CDT1, HIST4H4, NDC80, RFC2, TIPIN, BCL6, MYCN, ATRX, PMF1, MIS12, CENPE, HMGN1, POLG2, CENPA, PURA, PIF1, ING2, ZBED1, NPM2, ZNF330, H2AFY, MAD2L1, KIF22, XRCC4, BUB3, BUB1, SYCE1, CBX1, SMARCA5, RB1, TUBG1, INCENP, CLASP1, JUNB, TTN, HDAC8, RAD51, RPA3, NCAPD2, SMARCE1, KLHDC3, H2AFY2, H1FNT, HMGB1, CHEK1, CDCA5, SUMO3, UPF1, ERCC4, TERF2, RPA4, RGS12, APTX, CLIP1, JUND, UBE2I, LRPPRC, FOXC1, BIRC5, JUN, DCTN2, TINF2, PURB, PAFAH1B1, MAF, ACD, TERF2IP, TNP1, PSEN1, TOP1, NOL6, NUFIP1, ERCC1, PAM, ZW10, LIG4, RAN, PSEN2, CHMP1A, SUGT1, MKI67IP
Nuclear chromosome	TIMELESS, STAG3, RPA2, MCM3, HMGB2, SMC2, POLA1, CBX5, RPA1, DMC1, MCM7, ZNF238, SUV39H1, SMC1A, ZMIZ2, SYCE2, RCC1, REPIN1, TIPIN, ATRX, PURA, PIF1, ZBED1, NPM2, H2AFY, SYCE1, CBX1, TUBG1, TTN, HDAC8, RAD51, RPA3, SMARCE1, KLHDC3, H2AFY2, H1FNT, CHEK1, ERCC4, RPA4, RGS12, UBE2I, LRPPRC, FOXC1, JUN, PURB, ACD, TERF2IP, NOL6, NUFIP1, ERCC1, PAM, CHMP1A, MKI67IP
Nuclear chromosome part	TIMELESS, STAG3, RPA2, MCM3, POLA1, CBX5, RPA1, MCM7, ZMIZ2, SYCE2, RCC1, REPIN1, TIPIN, ATRX, PURA, PIF1, NPM2, H2AFY, SYCE1, CBX1, RPA3, KLHDC3, H2AFY2, H1FNT, ERCC4, RPA4, UBE2I, FOXC1, PURB, ACD, NUFIP1, ERCC1, PAM
Condensed chromosome	STAG3, SMC4, HMGB2, SMC2, NSL1, CENPF, DMC1, SUV39H1, SMC1A, DSN1, SYCE2, RCC1, PMF1, MIS12, CENPE, XRCC4, SYCE1, SMARCA5, TUBG1, TTN, RAD51, NCAPD2, HMGB1, CHEK1, RGS12, UBE2I, LRPPRC, NOL6, PAM, LIG4, CHMP1A, MKI67IP

Chromosomepericentric region	NSL1, ZWINT, CENPF, HELLS, ZWILCH, BUB1B, CENPC1, SMC1A, DSN1, APC, NDC80, PMF1, MIS12, CENPE, CENPA, ZNF330, MAD2L1, KIF22, BUB3, BUB1, INCENP, CLASP1, SUMO3, CLIP1, BIRC5, DCTN2, PAFAH1B1, PSEN1, ZW10, PSEN2, SUGT1
Immunological synapse	SYK, TRAT1, LAT, ZAP70, CD247, CD79A, GZMB, GZMA, BCL10, CARD11, MYH9
Kegg Pathways	Genes
primary immunodeficiency	UNG, DCLRE1C, CD40LG, TNFRSF13C, RFX5, LCK, CD8B, TNFRSF13B, CD3E, ZAP70, AICDA, BTK, IL2RG, CIITA, PTPRC, JAK3, CD19, CD79A, CD4, CD3D, CD8A, BLNK, IGLL1, AIRE, RFXAP, ICOS, CD40, RFXANK, TAP1, IKBKG, IL7R, TAP2, RAG2, RAG1, ADA
Intestinal immune network for IGA production	MAP3K14, CD40LG, TNFRSF13C, CD28, TNFRSF13B, AICDA, CXCR4, TNFRSF17, IL2, HLA-DMB, HLA-DOA, IL15, ITGB7, ICOSLG, TNFSF13, ITGA4, HLA-DQA2, HLA-DPB1, HLA-DRA, HLA-DPA1, HLA-DQA1, IL4, CCL25, CCR9, HLA-DMA, CCL28, ICOS, HLA-DOB, HLA-DQB1, HLA-DRB5, CD40, HLA-DRB1, CXCL12, CCR10, CD80, IL15RA, TNFSF13B, IL5, MADCAM1, CD86, PIGR, IL10, IL6, CCL27, TGFB1, LTBR
DNA replication	MCM2, RPA2, MCM5, MCM6, MCM3, RFC5, LIG1, PRIM1, POLA1, PCNA, POLE, DNA2, RPA1, RFC4, POLE2, MCM7, RFC3, RFC1, POLD1, POLD3, RNASEH2A, MCM4, POLE3, POLA2, RFC2, FEN1, PRIM2, RNASEH2B, RPA3, RPA4, RNASEH2C, POLE4, RNASEH1, POLD2, SSBP1, POLD4
mismatch repair	RPA2, RFC5, LIG1, MSH6, MSH2, MSH3, PCNA, MLH1, RPA1, RFC4, RFC3, RFC1, POLD1, POLD3, EXO1, RFC2, RPA3, MLH3, RPA4, POLD2, PMS2, SSBP1, POLD4

Table 8. *List of genes bearing somatic mutations in immune-related pathways significantly over-represented in the infected group in HNSC.*

Pathway	Genes
Lymphocyte activation	CD38, ITGAL, BTK, FYN, BTN3A1, POU2F2, HDAC9, RC3H2, KLF6, PRKCZ, GAL, RORA, JMJD6, DLG1, MLH1, PAG1, IL4R, APBB1IP, ITCH, TCF7, MEF2C, CYLD, IGBP1, ICAM1, TYRO3, MSH2, BLNK, IL12RB1, LGALS1, MYH9, EP300, CD40, NFATC2, CHRNA4, ATP11C, BMX, ELF4, PYCARD, IL21R, RELB, JAK3, PIK3CG, LFNG, EPHB6, AHR, GLI3, DOCK8, GATA3, PPP3CB, MAP3K8, CCL2, PRKAR1A, SUPT6H, WHSC1, ZBTB16, PTPN6, AICDA, TREML2, DUSP22, BCL6, SOS1, LEPR, CD46, STK11, MYB, GPAM, HELLS, IFNA8, HSPH1, EGR1, KIAA0922, CD80, CCR2, PIK3CA, LAX1, NCKAP1L, PREX1, IRF1, PCID2, EIF2AK4, RIPK3, EPO, LILRB2, FLOT2, AP3B1, NOTCH2, DOCK2, HAVCR2, DTX1, MAP3K7, LCP1, RAC1, MARCH7, IL10, CD27, ERBB2, AKT1, HSPD1, PIK3R1, CDKN2A, GSN, ATM, ITGB1, BANK1, MERTK, EPHB1, BATF, PAXIP1, BRAF, NCK1, RLTPR, PKNOX1, ITGB2, PGLYRP2, IL23R, NCSTN, AZI2, ICOS, RICTOR, RAG1, CLEC4E, B2M, AP1G1, AXL, MLST8, NLRC3, FADD, INPP5D, APLF, SOCS5, CHD7, PIK3CD, MALT1, EXO1, RAG2, PTPN2, TBC1D10C, FZD8, PAK2, PRF1, SATB1, JAG2, IFNE, ZFP36L1, IRS2, NHEJ1, BLOC1S3, NCR1, SEMA4A, DLL1, MAFB, HLA-DOA, PSMB10, LAT, FNIP1, HLA-DMB, PRKDC
Fc gamma R-mediated phagocytosis	PIK3CB, PLD1, AMPH, DNM2, PTPRC, MAPK1, PIK3CG, LIMK1, PIP5K1B, RPS6KB1, PLA2G4A, AKT3, PIK3CA, PLD2, RAF1, WASF3, DOCK2, RAC1, PIK3R5, VAV1, AKT1, FCGR2A, PIP5K1A, PIK3R1, GSN, FCGR1A, ASAP2, ASAP1, PRKCA, WASF2, VAV2, INPPL1, PLA2G4F, INPP5D, MAP2K1, PIK3CD, CFL1, MARCKSL1, RPS6KB2, LIMK2, PIP5K1C, PLA2G4E, LAT

Table 9. *Count and ratio (number of mutations found divided by the number of sample) of somatic mutations present in 279 samples in HNSC.*

	Samples	Infection (%)	Frame Shift Del.	Frame Shift Del. (ratio)	Frame Shift Ins.	Frame Shift Ins. (ratio)	In Frame Del.	In Frame Del. (ratio)
Inf.	39	13.98	131.00	3.36	55.00	1.41	38.00	0.97
Non-Inf.	240	86.02	1039.00	4.33	474.00	1.98	245.00	1.02
Total	279	100.00	1170.00	4.19	529.00	1.90	283.00	1.01

	Samples	In Frame Ins.	In Frame Ins. (ratio)	Missense Mutation	Missense Mutation (ratio)	Nonsense Mutation	Nonsense Mutation (ratio)
Inf.	39.00	4.00	0.10	4540.00	116.41	362.00	9.28
Non-Inf.	240.00	37.00	0.15	28720.00	119.67	2324.00	9.68
Total	279.00	41.00	0.15	33260.00	119.21	2686.00	9.63

	Samples	Nonstop Mutation	Nonstop Mutation (ratio)	Silent	Silent (ratio)	Splice Site	Splice Site (ratio)
Inf.	39.00	9.00	0.23	1742.00	44.67	77.00	1.97
Non-Inf.	240.00	35.00	0.15	11180.00	46.58	787.00	3.28
Total	279.00	44.00	0.16	12922.00	46.32	864.00	3.10

Table 10. *Function of coding genes in HPV16 virus*

Gene	Function
E6	Oncoprotein. It has long been recognized as a potent oncogene and is intimately associated with the events that result in the malignant conversion of virally infected cells.
E7	Plays a role in viral genome replication by driving entry of quiescent cells into the cell cycle.
E1	The Papillomaviriuses are icosahedral dsDNA viruses that encode 7 early proteins, E1 to E7, and 2 late structural proteins L1 and L2. Host infection by members of the papillomaviruses gives rise to warts and cancer in some instances. The E1 helicase protein is an ATP-dependent DNA helicase that facilitates the inititiation of viral DNA replication through its interactions with the viral E2 protein. The E1-E2 complex binds to the viral origin of replication where E1 acts to unwind the DNA and E2 acts as a transcriptional activator.
E2	Regulatory protein.
E4	Although located in the early part of the viral genome, the E4 proteins are primarily expressed during the late stages of infection, at or around the time that genome amplification is initiated. E4 proteins have been shown to inhibit cell proliferation in G2, and to participate in efficient genome amplification.
E5	Oncoprotein. Contributes to cell transformation. Produces polyploid cells by endoreplication.
L2	Minor capsid protein.
L1	Major capsid L1 protein; Two structural proteins are involved in papillomavirus capsid formation, a major (L1) and a minor (L2) protein; L1 forms the pentameric assembly unit of the viral shell while L2 mediates several facets of viral entry including endosomal escape after uncoating.

Table 11. *Expression of HPV16 genes when infecting CESC and HNSC, obtained through HTSeq. Number of reads aligned in each coding sequences in HPV16. Number of reads with no feature represents the number of reads that could not align completely with any feature. Ambiguous reads are the ones which have been allocated in more than one feature. Too low aQual represent the reads with alignment quality below 10 (by default). Not aligned reads are reads without alignment in the SAM file. Finally, reads in alignment not unique are the ones which have more than one alignment.*

			E6	E7	E1	E2	E4	E5	L2	L1	Total aligned reads	No feature	Ambiguous	Too low aQual	Not aligned	Alignment not unique
CESC	TCGA-ZJ-AAXD-01A-21R-A42T-07	Counts	1534	1548	772	0	1825	592	43	107	3339	11598	3395	155	1420079	0
		Percentage	45.94%	46.36%	23.12%	0.00%	54.66%	17.73%	1.29%	3.20%	100.00%	–	–	–	–	–
	TCGA-VS-A9V5-01A-11R-A42T-07	Counts	4951	6445	9099	0	5311	2336	626	1264	18636	44924	9472	313	2341402	0
		Percentage	26.57%	34.58%	48.82%	0.00%	28.50%	12.53%	3.36%	6.78%	100.00%	–	–	–	–	–
	TCGA-VS-A9UD-01A-11R-A42T-07	Counts	299	500	772	0	2983	2509	124	5188	11576	16657	2752	355	1724084	0
		Percentage	2.58%	4.32%	6.67%	0.00%	25.77%	21.67%	1.07%	44.82%	100.00%	–	–	–	–	–
	TCGA-Q1-A6DT-01A-11R-A32P-07	Counts	1339	1763	5669	0	18515	7511	519	1182	33396	72376	25492	1065	3182368	0
		Percentage	4.01%	5.28%	16.98%	0.00%	55.44%	22.49%	1.55%	3.54%	100.00%	–	–	–	–	–
	TCGA-EA-A97N-01A-11R-A38B-07	Counts	2092	3326	1793	0	2384	865	45	83	5170	17082	4692	447	2571472	0
		Percentage	40.46%	64.33%	34.68%	0.00%	46.11%	16.73%	0.87%	1.61%	100.00%	–	–	–	–	–
HNSC	TCGA-QK-A6IF-01A-11R-A31N-07	Counts	1278	1589	2595	0	154	0	0	4	2753	5096	670	0	1941273	0
		Percentage	46.42%	57.72%	94.26%	0.00%	5.59%	0.00%	0.00%	0.15%	100.00%	–	–	–	–	–
	TCGA-HD-A634-01A-11R-A28V-07	Counts	2573	3120	1842	0	6343	4237	23945	1927	38294	54039	9376	0	3522307	0
		Percentage	6.72%	8.15%	4.81%	0.00%	16.56%	11.06%	62.53%	5.03%	100.00%	–	–	–	–	–
	TCGA-CR-7385-01A-11R-2016-07	Counts	3010	3700	4068	0	2068	763	111	73	7083	17919	4501	0	2596620	0
		Percentage	42.50%	52.24%	57.43%	0.00%	29.20%	10.77%	1.57%	1.03%	100.00%	–	–	–	–	–
	TCGA-BB-7866-01A-11R-2232-07	Counts	7206	7736	7571	0	1725	2526	2726	6715	21263	42832	5666	0	3174764	0
		Percentage	33.89%	36.38%	35.61%	0.00%	8.11%	11.88%	12.82%	31.58%	100.00%	–	–	–	–	–
	TCGA-BA-A4IH-01A-11R-A266-07	Counts	5482	4476	4507	0	5341	1772	395	5174	17189	35755	8216	0	4690191	0
		Percentage	31.89%	26.04%	26.22%	0.00%	31.07%	10.31%	2.30%	30.10%	100.00%	–	–	–	–	–

Table 12. *Product of coding sequences in HBV virus*

Coding sequences	Product	Function
CDS0	Polymerase	
CDS1	Large envelope protein	
CDS2	Middle envelope protein	
CDS3	Small envelope protein	
CDS4	X protein	Promotes cell cycle progression and inhibits tumor suppressor protein
CDS5	Pre-capsid protein	
CDS6	Capsid protein	

Table 13. *Expression of HBV genes when infecting LIHC samples, obtained through HTSeq. Number of reads aligned in each coding sequences inHPV16. Number of reads with no feature represents the number of reads that could not align completely with any feature. Ambiguous reads are the ones which have been allocated in more than one feature. Too low aQual represent the reads with alignment quality below 10 (by default). Not aligned reads are reads without alignment in the SAM file. Finally, reads in alignment not unique are the ones which have one more than one alignment.*

			cds0	cds1	cds2	cds3	cds4	cds5	cds6	Total aligned reads	No feature	Ambiguous	Too low aQual	Not aligned	Alignment not unique
LIHC	TCGA-2Y-A9H4	Count	1	0	0	4765	8908	38	0	13712	19985	129	0	2965024	0
		Percentage	0.01%	0%	0%	34.75%	64.96%	0.28%	0%	100.00%	–	–	–	–	–
	TCGA-CC-A7IK	Count	371	0	0	4324	4272	363	0	9330	15724	1975	0	4236928	0
		Percentage	3.98%	0%	0%	46.35%	45.79%	3.89%	0%	100.00%	–	–	–	–	–
	TCGA-DD-AAEK	Count	52	0	0	44031	25638	313	0	70034	123797	722	0	2273311	0
		Percentage	0.07%	0%	0%	62.87%	36.61%	0.45%	0%	100.00%	–	–	–	–	–
	TCGA-ED-A7XP	Count	21	0	0	16861	21371	56	0	38309	60247	337	0	3854298	0
		Percentage	0.05%	0%	0%	44.01%	55.79%	0.15%	0%	100.00%	–	–	–	–	–
	TCGA-G3-A25U	Count	14	0	0	29449	22914	0	0	52377	92052	26	0	3572354	0
		Percentage	0.03%	0%	0%	56.23%	43.75%	0%	0%	100.00%	–	–	–	–	–